Agilent G2721AA
Spectrum Mill MS
Proteomics Workbench

Familiarization Guide

Agilent Technologies

Notices

© Agilent Technologies, Inc. 2007

No part of this manual may be reproduced in any form or by any means (including electronic storage and retrieval or translation into a foreign language) without prior agreement and written consent from Agilent Technologies, Inc. as governed by United States and international copyright laws.

Manual Part Number

G2721-90032

Edition

Fifth edition, June 2007

Printed in USA

Agilent Technologies, Inc.
5301 Stevens Creek Blvd.
Santa Clara, CA USA 95052

If you have any comments about this guide, please send an e-mail to feedback_lcms@agilent.com.

Warranty

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end user customers. Agilent provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

In this Guide...

The *Familiarization Guide* presents step-by-step exercises to analyze example data files with the Spectrum Mill MS Proteomics Workbench. In this guide you process protein digest data to learn:

- How to extract spectra from MS/MS and MS-only data sets
- How to perform protein database searches
- How to review and validate search results
- How to perform *de novo* sequencing
- How to process data from Agilent ion traps (CID and ETD), Agilent Q-TOF, and Agilent TOF
- How to process data from labeling studies such as ICAT experiments
- How to process data from iTRAQ experiments

If you have any comments about this guide, please send an e-mail to feedback_lcms@agilent.com.

1 First Steps for Processing MS/MS Data

In these exercises, you process and compare data from two *E. coli* cell states. These exercises illustrate the basic processing steps you use for many types of samples.

2 Optional Processing for MS/MS Data

In these exercises, you process the *E. coli* data with advanced features, such as variable modifications search and *de novo* sequencing.

3 Processing Ion Trap - ETD Data

These exercises show how to process electron transfer dissociation (ETD) data from an Agilent ion trap.

4 Processing Q-TOF MS/MS Data

In these exercises, you process an example Agilent Q-TOF MS/MS data file. Because this sample originated with the Agilent OFFGEL Fractionator, you also incorporate peptide pI into the autovalidation scheme.

5 Searching for Unknown Modifications

These exercises show how to search MS/MS spectra to identify unsuspected modifications. The chapter illustrates this “mass gap” search with an Agilent Q-TOF data file.

6 Processing ICAT Data

In these exercises, you apply the Spectrum Mill processing to an ICAT sample.

7 Processing iTRAQ Data

In these exercises, you apply the Spectrum Mill processing to an iTRAQ sample.

8 First Steps for Processing MS-Only Data

These exercises show how to process LC-ESI-TOF and AP-MALDI-TOF MS-only (PMF) data.

Contents

1 First Steps for Processing MS/MS Data 9

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 11
Exercise 2. Access the Spectrum Mill server 14
Exercise 3. Extract the spectra 16
Exercise 4. Create a species subset database 22
Exercise 5. Search the database 24
Exercise 6. Autovalidate the hiqhest-quality results 30
Exercise 7. (Optional) Examine the validated results 33
Exercise 8. Display medium-quality results in preparation for manual validation 37
Exercise 9. Manually review and validate results 42
Exercise 10. Examine results with the Spectrum Viewer 44
Exercise 11. Summarize valid results 47
Exercise 12. Compare results from two samples 51
Exercise 13. (Optional) Examine details of search results 55

2 Optional Processing for MS/MS Data 61

Exercise 1. Create a list of identified proteins for variable modifications search 62
Exercise 2. Search in variable modifications mode 64
Exercise 3. Review and validate variable modifications search results 69
Exercise 4. Check for remaining high-quality spectra 72
Exercise 5. Perform Sherenga *de novo* sequencing 77
Exercise 6. Review results from Sherenga *de novo* sequencing 80

3 Processing Ion Trap - ETD Data 85

Exercise 1. (Optional) Transfer ETD data file to the Spectrum Mill server 86
Exercise 2. Extract the spectra 88

Exercise 3. Create indices for a user database 92
Exercise 4. Run database search 94
Exercise 5. Run Autovalidation 100
Exercise 6. Display valid database search results in peptide mode 103
Exercise 7. Examine results with the Spectrum Viewer 106
Exercise 8. Display valid database search results in protein mode 109

4 Processing Q-TOF MS/MS Data 113

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 114
Exercise 2. Extract the spectra 116
Exercise 3. Search a database 121
Exercise 4. Autovalidate the hiqhest-quality results 125
Exercise 5. (Optional) Examine the validated results 129
Exercise 6. Display medium-quality results in preparation for manual validation 134
Exercise 7. Manually review and validate results 140
Exercise 8. Summarize valid results 142

5 Searching for Unknown Modifications 147

Exercise 1. Create a list of identified proteins 149
Exercise 2. Search for unknown modifications 151
Exercise 3. Autovalidate the highest-quality results 156
Exercise 4. Examine the list of validated peptides 158
Exercise 5. Examine details of search results to determine modification sites 163

6 Processing ICAT Data 167

Exercise 1. (Optional) Transfer ICAT data file to Spectrum Mill server 169
Exercise 2. (Optional) Extract spectra from ICAT data 171
Exercise 3. Create indices for a user database 176
Exercise 4. Run database search on ICAT data 178
Exercise 5. Autovalidate the ICAT data 183
Exercise 6. Summarize valid results in peptide mode 186
Exercise 7. Summarize valid results in protein mode 190

7 Processing iTRAQ Data 193

Exercise 1. (Optional) Transfer iTRAQ data file to Spectrum Mill server 194
Exercise 2. (Optional) Extract the spectra 196
Exercise 3. Search a database 201
Exercise 4. Autovalidate the hiqhest-quality results 206
Exercise 5. Create a file of iTRAQ correction factors 208
Exercise 6. Apply the iTRAQ correction factors to your sample 210
Exercise 7. Summarize valid results in peptide mode 212
Exercise 8. Summarize valid results in protein mode 216

8 First Steps for Processing MS-Only Data 219

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 220
Exercise 2. Extract the LC-ESI-TOF data 222
Exercise 3. Create a species subset database 228
Exercise 4. Search the LC-ESI-TOF data 230
Exercise 5. Display PMF Search results for LC-ESI-TOF data 234
Exercise 6. Extract the AP-MALDI-TOF data 236
Exercise 7. Search the AP-MALDI-TOF data 240
Exercise 8. Display PMF Search results for AP-MALDI-TOF data 243
Exercise 9. Review AP-MALDI-TOF PMF Search results 246

Contents

1
First Steps for Processing MS/MS Data

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 11
Exercise 2. Access the Spectrum Mill server 14
Exercise 3. Extract the spectra 16
Exercise 4. Create a species subset database 22
Exercise 5. Search the database 24
Exercise 6. Autovalidate the hiqhest-quality results 30
Exercise 7. (Optional) Examine the validated results 33
Exercise 8. Display medium-quality results in preparation for manual validation 37
Exercise 9. Manually review and validate results 42
Exercise 10. Examine results with the Spectrum Viewer 44
Exercise 11. Summarize valid results 47
Exercise 12. Compare results from two samples 51
Exercise 13. (Optional) Examine details of search results 55

This chapter provides familiarization exercises to process an example set of MS/MS data files with the Agilent Spectrum Mill MS Proteomics Workbench. This material elaborates on the first section of Chapter 1 of the companion *Spectrum Mill MS Proteomics Workbench Application Guide*. This chapter is intended as an introduction to the basic operation of the software; you are encouraged to customize these steps for your samples and studies.

The exercises here assume that the Spectrum Mill workbench has already been installed on your server, the **NCBInr** database has been downloaded and indexed, and the system is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

While the exercises in this chapter use Agilent ion trap data files, the general workflow applies broadly for MS/MS data files. However, the specific settings, such as mass tolerance, may vary for MS/MS data from other types of instruments. This guide additionally includes exercises that are specific to other MS/MS data types:

- If your interest is ion trap data from electron transfer dissociation (ETD), please see Chapter 3, “Processing Ion Trap - ETD Data,” starting on page 85.
- If you have Agilent Q-TOF data, please see Chapter 4, “Processing Q-TOF MS/MS Data,” starting on page 113.

After you have worked through the exercises in this chapter and have a good understanding of the software, we recommend you try the Easy MS/MS Search form. This form streamlines the data processing steps by consolidating the Data Extractor, MS/MS Search, Autovalidation, and Protein/Peptide Summary pages onto a single form that shows only the most important parameters. For more details on Easy MS/MS Search, see the online help for that form.

Exercise 1. (Optional) Transfer data files to Spectrum Mill server

Data transfer to the Spectrum Mill server is straightforward. For data from most instruments, you merely move or copy the files into the appropriate directory on the Spectrum Mill server, without any other preparation. This is the case for the following types of data files:

- Agilent ion trap and Q-TOF *.d
- Thermo Fisher Scientific LCQ *.raw
- Applied Biosystems/MDS Sciex QSTAR *.wiff

For Micromass Q-Tof data, you first preprocess the data files with the Micromass data system, and then you copy or move the processed spectral files to the server.

This exercise serves as an example for the raw data file types (*.d, *. raw, or *.wiff).

To make it easy to compare data sets, you must set up the appropriate directory structure for your files on the Spectrum Mill server. Whenever you want to compare samples in a set, you need to set up a subdirectory for each sample. This subdirectory may contain data files from multiple sample fractions or gel slices. In this exercise, you set up subdirectories for two types of *E. coli* samples — a stressed sample and a control sample. Into each subdirectory, you copy or move data files from the two-dimensional (2D) LC/MS/MS analysis of sample fractions.

NOTE The example data files may have been installed on the server at the time the Spectrum Mill workbench was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

Steps	Detailed Instructions	Comments
1 Copy the example data file directories from the **X:\ExampleData\msdataSM\ExampleData\Agilent\Trap\ecoli** directory on the *Spectrum Mill Example Data CD* onto your client PC.	• Copy the subdirectories **heatshock** and **wild** from the *Spectrum Mill Example Data CD* to any directory on your client PC.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • This exercise simulates the likely laboratory scenario of file transfer from a client or instrument PC to the server; you may transfer files directly from the CD to the server if you prefer.
2 Create the following folders for the example files on the Spectrum Mill server: ◦ **msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock** ◦ **msdataSM\ExampleData\Agilent\Trap\ecoli\wild**	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In **msdataSM**, create a new system of folders for the example files. You need two new sample folders because there are two *E. coli* samples — heatshock (stressed) and wild (control). Each has multiple 2D LC/MS/MS fractions: ◦ Create **msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock**. ◦ Create **msdataSM\ExampleData\Agilent\Trap\ecoli\wild**. ◦ To create a new folder, click **File > New> Folder**, and type in the name of the folder.	• If you do not know how to find your Spectrum Mill file system, ask the person who installed your software. • **Do not include spaces or parentheses in your directory names.** • When you process your own samples, remember to set up a separate folder for each sample. Each folder should contain all sample fractions. • You may create up to ten folders between **msdataSM** and your data files, but shorter path lengths reduce memory usage, especially for large data sets.
3 Copy the example files to the new directory.	• Copy or move the **heatshock** and **wild** example raw files from your client PC to the new folders on the server PC.	When you process your own data, remove any spaces or parentheses in the data file names. For example, change the file **my sample.d** to **my_sample.d**

Steps	Detailed Instructions	Comments
4 Make sure you have both read and write permissions for the data folders you just created.	**a** Right-click the **heatshock** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked, and apply the changes to all subfolders and files. **c** Click the **Security** tab. **d** Make sure all user groups have full permissions. **e** Click **OK**. **f** Repeat steps **a** through **e** for the **wild** folder.	If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors.

Exercise 2. Access the Spectrum Mill server

Steps	Detailed Instructions	Comments
1 Start the Spectrum Mill workbench on your PC.	• Double-click the desktop icon to launch the Spectrum Mill workbench.	If you don't have this icon, go to the next step.
2 If you don't see the Spectrum Mill icon, launch your web browser and type the URL for the Spectrum Mill home page.	**a** Click your Internet Explorer icon. **b** In the web browser window, type the URL for the Spectrum Mill server.	If you don't know the URL, ask the person who installed the software.
3 Check that you see the Spectrum Mill home page.	• Check that you see the page shown in Figure 1 on page 15.	
4 If you do not, check that the server is booted up and that your URL is correct.		
5 (Optional) Make this your home page.	• See instructions in your web browser help. Look for "home page."	If you make this your home page, you can quickly navigate back to the Spectrum Mill home page with a single click of the **home** icon.
6 Close the web browser window.		

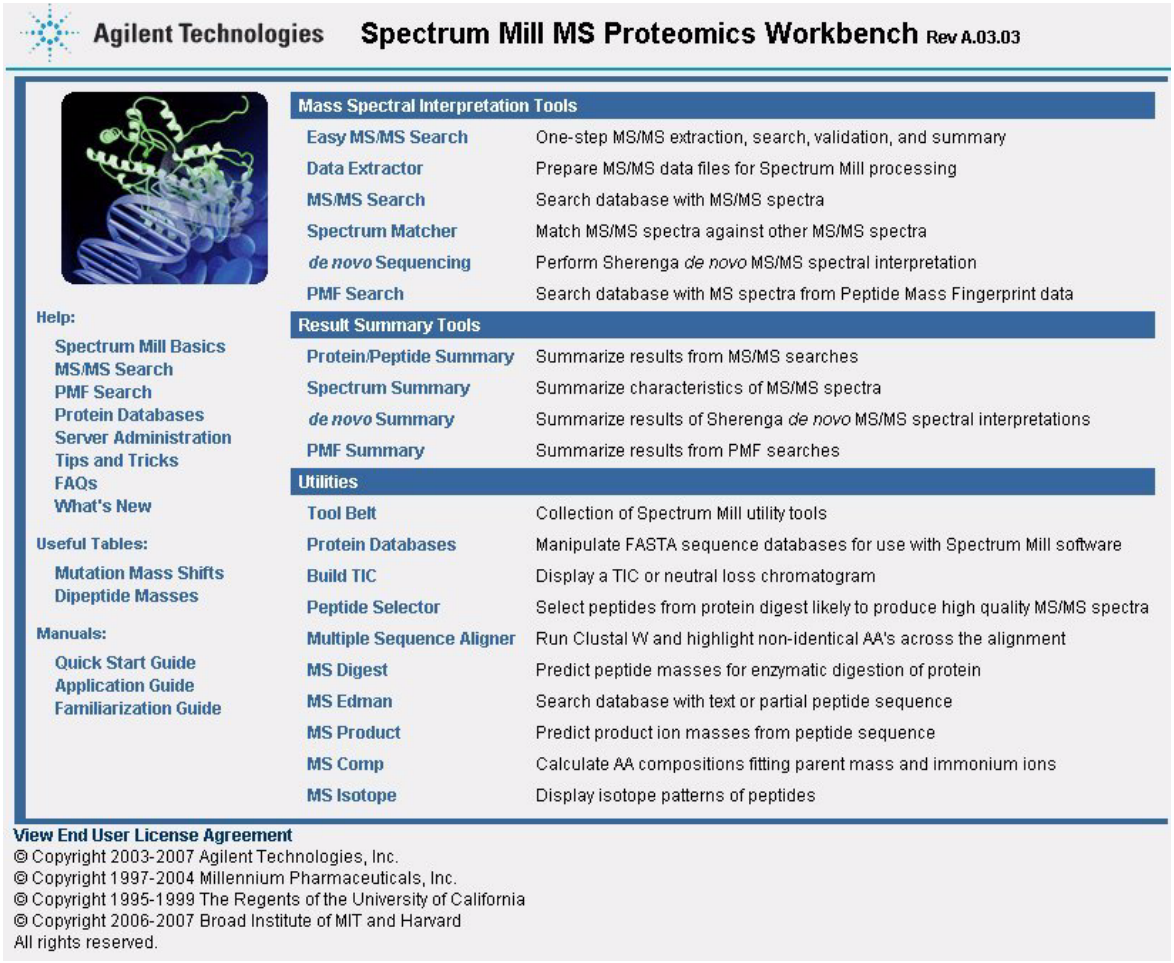
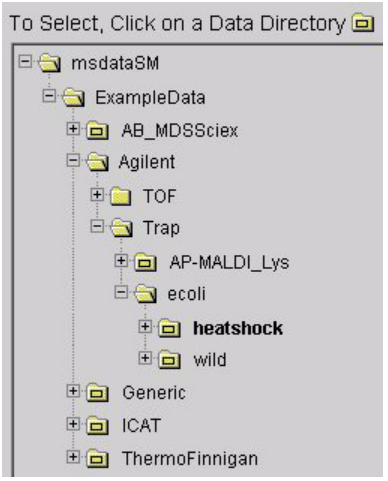

Agilent Technologies
Spectrum Mill MS Proteomics Workbench Rev A.03.03
Help:
Spectrum Mill Basics
MS/MS Search
PMF Search
Protein Databases
Server Administration
Tips and Tricks
FAQs
What's New
Useful Tables:
Mutation Mass Shifts
Dipeptide Masses
Manuals:
Quick Start Guide
Application Guide
Familiarization Guide
Mass Spectral Interpretation Tools
Easy MS/MS Search
One-step MS/MS extraction, search, validation, and summary
Data Extractor
Prepare MS/MS data files for Spectrum Mill processing
MS/MS Search
Search database with MS/MS spectra
Spectrum Matcher
Match MS/MS spectra against other MS/MS spectra
de novo Sequencing
Perform Sherenga de novo MS/MS spectral interpretation
PMF Search
Search database with MS spectra from Peptide Mass Fingerprint data
Result Summary Tools
Protein/Peptide Summary
Summarize results from MS/MS searches
Spectrum Summary
Summarize characteristics of MS/MS spectra
de novo Summary
Summarize results of Sherenga de novo MS/MS spectral interpretations
PMF Summary
Summarize results from PMF searches
Utilities
Tool Belt
Collection of Spectrum Mill utility tools
Protein Databases
Manipulate FASTA sequence databases for use with Spectrum Mill software
Build TIC
Display a TIC or neutral loss chromatogram
Peptide Selector
Select peptides from protein digest likely to produce high quality MS/MS spectra
Multiple Sequence Aligner
Run Clustal W and highlight non-identical AA's across the alignment
MS Digest
Predict peptide masses for enzymatic digestion of protein
MS Edman
Search database with text or partial peptide sequence
MS Product
Predict product ion masses from peptide sequence
MS Comp
Calculate AA compositions fitting parent mass and immonium ions
MS Isotope
Display isotope patterns of peptides
View End User License Agreement
© Copyright 2003-2007 Agilent Technologies, Inc.
© Copyright 1997-2004 Millennium Pharmaceuticals, Inc.
© Copyright 1995-1999 The Regents of the University of California
© Copyright 2006-2007 Broad Institute of MIT and Harvard
All rights reserved.

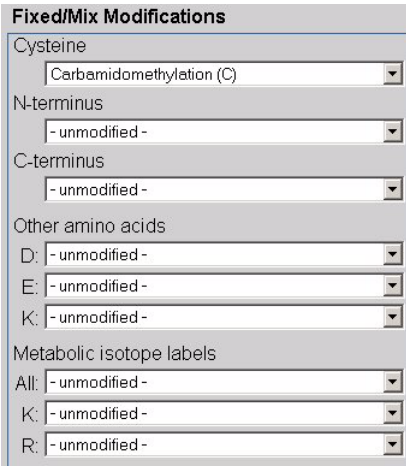
Figure 1 Spectrum Mill home page

Exercise 3. Extract the spectra

The Spectrum Mill Data Extractor preprocesses raw data files to extract high-quality spectra for database searches. The Data Extractor automatically detects which type of raw file you have and then invokes the appropriate extraction program (provided that it has been purchased and installed on your server). The MS/MS raw file data extractors extract and merge nearby MS/MS spectra from the same precursor ion. They optionally apply MS/MS similarity criteria prior to merging scans, to avoid merging closely eluting or co-eluting isobaric peptides. For Agilent *.d ion trap and Thermo Fisher Scientific *.raw ion trap data, the extractors optionally merge MS^2 and MS^3 scans from the same precursor. The extractors assign precursor charges where possible, centroid the MS/MS spectra, calculate spectral features, filter MS/MS spectra by quality, and calculate extracted ion chromatograms (EICs) for the intervening MS precursor scans. The latter are used for quantitation.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link.	

Steps	Detailed Instructions	Comments
2 Select the **msdataSM\ ExampleData\Agilent\Trap\ ecoli\heatshock** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **heatshock** to select that data file directory.  **c** Make sure that the name of the **heatshock** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 2 on page 20.	• Directories are identified by different types of icons: ◦ Plain folders indicate directories that do not have data files directly beneath them. ◦ Folders with rectangles indicate data directories. ◦ Folders with line spectra (bar graphs) indicate data files (which for Agilent files are also directories in themselves). • Try clicking the names of each type to see which turn **bold**, indicating that they are selectable. • The software remembers your data file selection when you go to other Spectrum Mill forms. • If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Click **OK.** The name of the modification appears as in Figure 2 on page 20.	• Carbamidomethylation is the default when you first install the Spectrum Mill workbench. • To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
4 Set parameters as shown in Figure 2 on page 20.	**a** Keep the defaults whenever appropriate. This example uses all default settings, except for **Data Directory**. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. You may also need to change the time range for **Merge scans with same precursor m/z** to be compatible with your chromatographic data. **c** Click a blue section divider bar to display help for that section of the form.	• You use the **Sequence tag length** to filter out noisy spectra. This is the longest sequence of amino acids that is represented by the fragments in a spectrum. • If you set this to **>1**, you ensure that all possible good spectra are extracted. You have opportunities to set more stringent requirements later when you perform the database search.

Steps	Detailed Instructions	Comments
5 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** area to the right of the Data Extractor form. See Figure 3 on page 21. **c** Scroll to the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.
6 (Optional) View the **cpick_in** subdirectory to see the files Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\Agilent\Trap\ecoli\heatshock\cpick_in**. **b** Notice the new files created there. Each one represents an extracted spectrum from your raw data file.	File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)

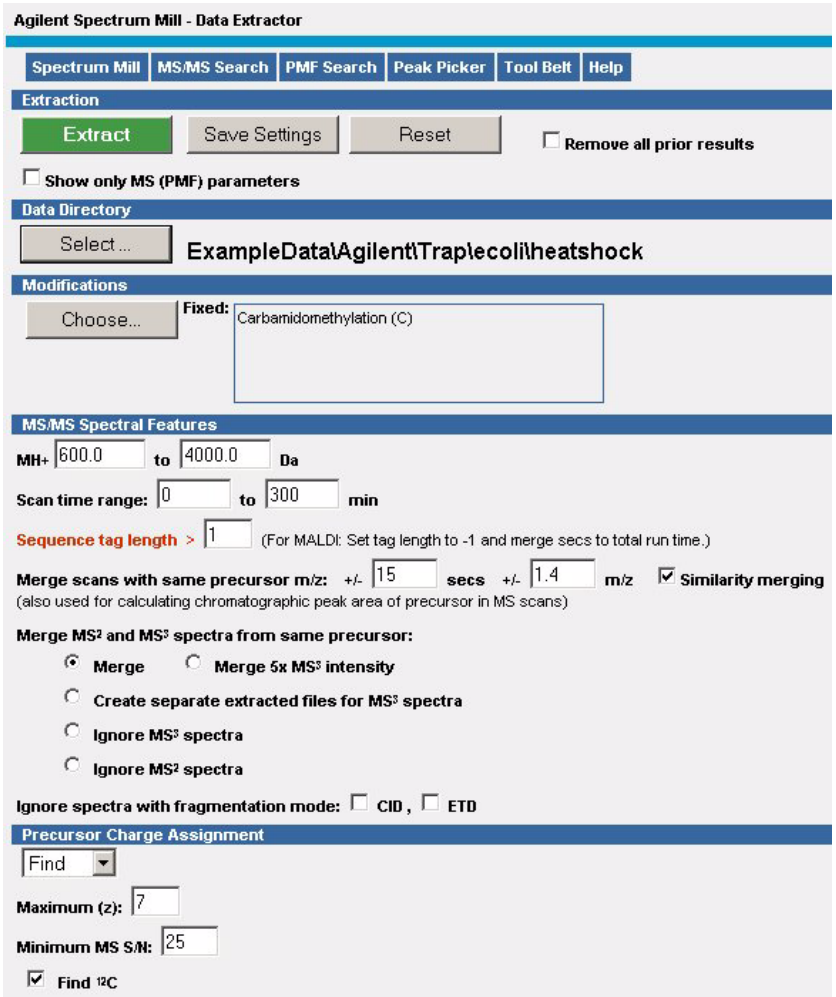


Figure 2 Settings for Data Extractor

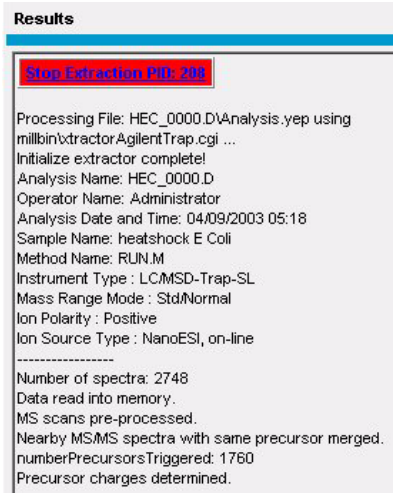
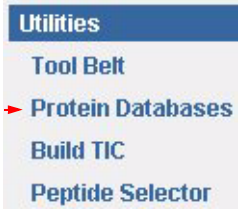


Figure 3 Data Extractor running

Exercise 4. Create a species subset database

The Spectrum Mill workbench allows you to create subset databases based on species and/or molecular weight filters. You can search these subset databases much more quickly than a full database.

Since you know the example data files contain primarily *E. coli* proteins, you create a subset *E. coli* species database.

Steps	Detailed Instructions	Comments
1 Make sure the public database **NCBInr** is available on your Spectrum Mill server.	• Ask your system administrator to install this database if that has not already been done.	To install a database, see the online help for **Protein Databases**. (The help document is **SpectrumMill\millhtml\SM_instruct\faman.htm**.
2 Navigate to the Protein Databases page.	• From the Spectrum Mill home page, click the Protein Databases link. 	
3 Open the form to create a species subset database.	• On the Protein Databases page, click the **Create species subset database** option.	See the form shown in Figure 4 on page 23.
4 Fill in the form as shown in Figure 4 on page 23.	• Check that your form looks similar to the one shown in Figure 4 on page 23.	
5 Start the subset creation.	**a** Click the **Create** button. **b** After a short wait, check the bottom of the page for text that indicates that the software is creating the subset database. See Figure 5 on page 23 as an example.	The number of database entries depends on when you downloaded the **NCBInr** database.

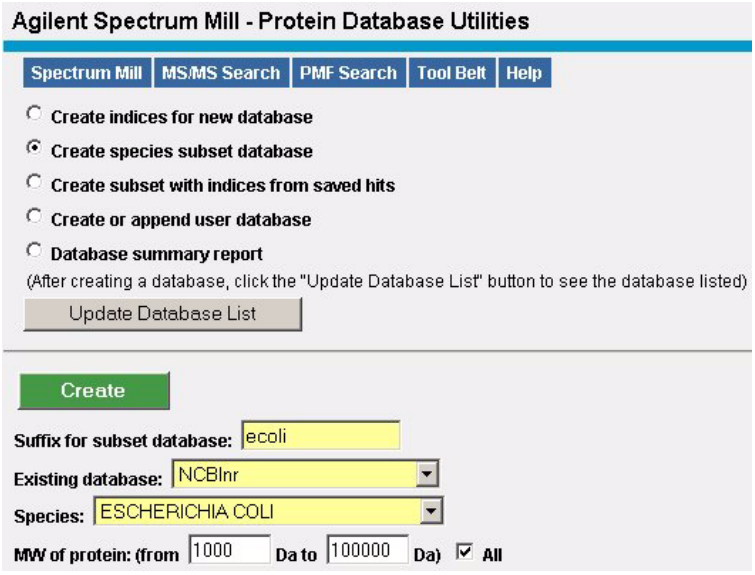

Agilent Spectrum Mill - Protein Database Utilities
Spectrum Mill
MS/MS Search
PMF Search
Tool Belt
Help
Create indices for new database
Create species subset database
Create subset with indices from saved hits
Create or append user database
Database summary report
(After creating a database, click the "Update Database List" button to see the database listed)
Update Database List
Create
Suffix for subset database: ecoli
Existing database: NCBInr
Species: ESCHERICHIA COLI
MW of protein: (from 1000 Da to 100000 Da) All

Figure 4 Form to create species subset database

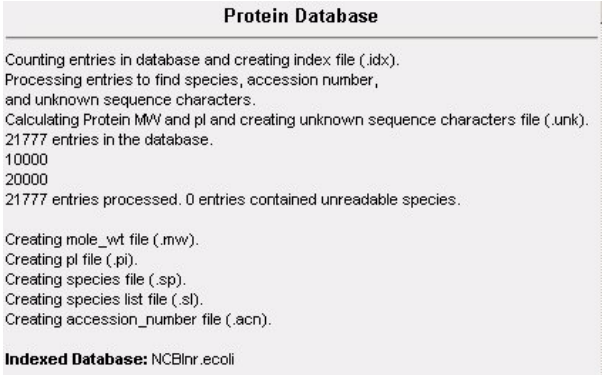
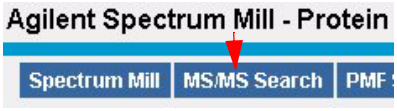

Protein Database
Counting entries in database and creating index file (.idx).
Processing entries to find species, accession number,
and unknown sequence characters.
Calculating Protein MW and pI and creating unknown sequence characters file (.unk).
21777 entries in the database.
10000
20000
21777 entries processed. 0 entries contained unreadable species.
Creating mole_wt file (.mw).
Creating pI file (.pi).
Creating species file (.sp).
Creating species list file (.sl).
Creating accession_number file (.acn).
Indexed Database: NCBInr.ecoli

Figure 5 Creation of species subset database

Exercise 5. Search the database

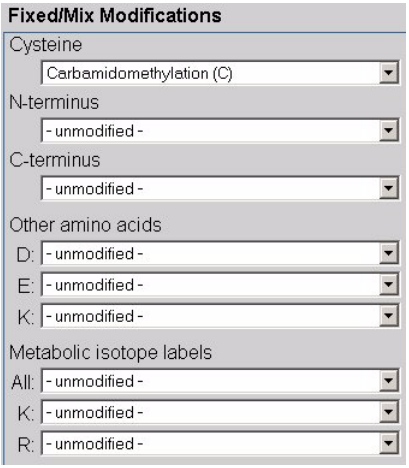
After you have extracted your spectra, you are ready to search each spectrum against a protein or DNA database. As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal of identifying as many spectra as possible. For this initial search, you use identity mode, which means the spectra must be consistent with the database sequence. For later searches, you may use variable modifications or homology mode, which means spectra may show modifications relative to the database sequence.

Steps	Detailed Instructions	Comments

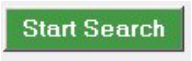
1 Navigate to the MS/MS Search page.	• Do one of the following:
	  ◦ From the Protein Databases page, click the **MS/MS Search** button.
	 
	  ◦ From the Spectrum Mill home page, click the link to MS/MS Search.
	
		Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Protein Databases page to the MS/MS Search page.

2 Check that your **Data Directory** is set to **heatshock**.	• If you have just performed the data extraction, your **heatshock** data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Select the *E. coli* subset database.	• For **Database** (under **Search Parameters**), select the name of the *E. coli* subset database you just created.	• Your *E. coli* subset database is likely named **NCBInr.ecoli** • If the database name does not appear, reload or refresh the MS/MS search form.

Steps	Detailed Instructions	Comments

4 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Click **OK.** The name of the modification appears as in Figure 6 on page 28.	• The modification is likely already set.
		• Three types of modifications are available for MS/MS Search:
		  ◦ The **fixed** modifications are assumed to apply universally and are searched in a single search cycle.
		  ◦ The **mix** modifications trigger cyclic MS/MS searches, where a different form of the modification is searched in each cycle.
		  ◦ The **variable** modifications allow for both modified and unmodified forms within a peptide. All variable modifications are searched in each cycle.
		• For more information about choosing modifications, see the online help.
		• Your system administrator can configure custom modifications.

Steps	Detailed Instructions	Comments
5 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Proton mobility scoring**. **c** Mark the check box for **Dynamic peak thresholding**. **d** Verify that the **Search mode** is set to **Identity**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Protein mobility scoring** is a scoring enhancement for ion trap data that applies a bonus or penalty based on the expected relative intensities for MS/MS fragments from a given peptide sequence. When you process your own ion trap data, it is best to use this scoring, except with modifications that significantly alter the fragmentation patterns of peptides. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.
6 Set other parameters as shown in Figure 6 on page 28.	**a** Keep the default settings. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	Changes from defaults are highlighted in yellow.

Steps	Detailed Instructions	Comments
7 Start the search.	**a** Click the **Start Search** button.  **b** View search progress in the **Results** area to the right of the MS/MS Search form. See Figure 7 on page 29. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • Search time varies depending on the size of the database searched. This search goes fairly quickly because you use a species subset database. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

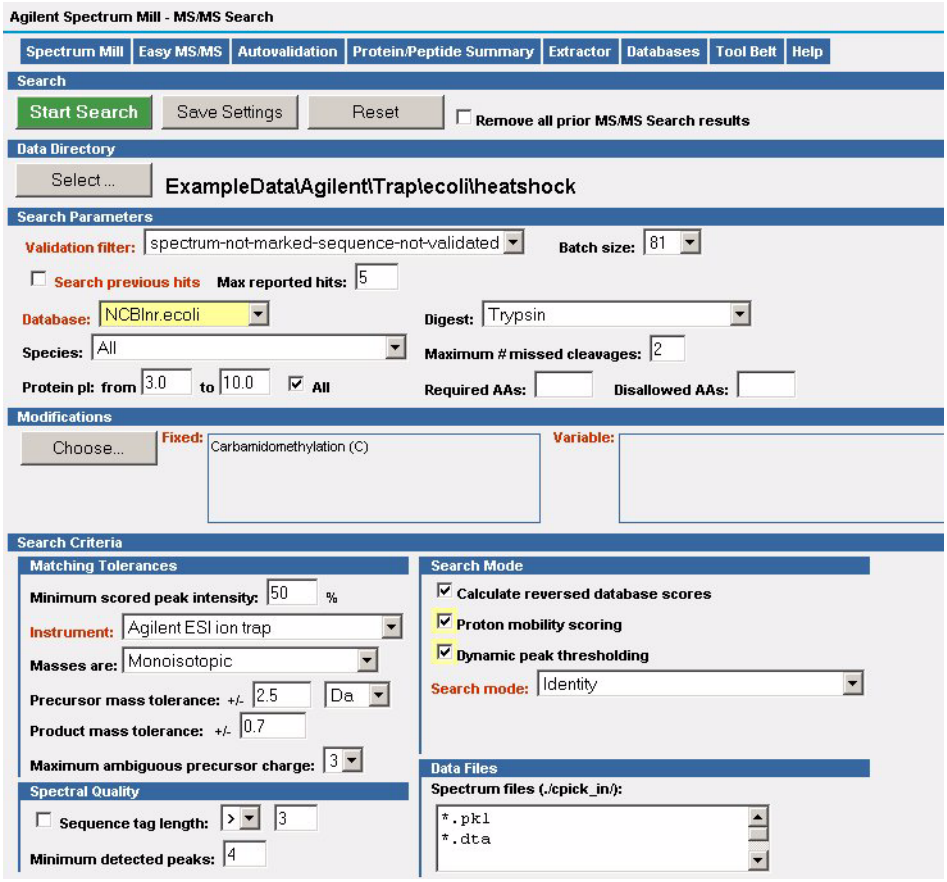
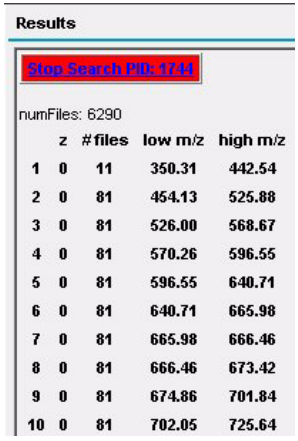


Figure 6 MS/MS Search settings



	z	# files	low m/z	high m/z
1	0	11	350.31	442.54
2	0	81	454.13	525.88
3	0	81	526.00	568.67
4	0	81	570.26	596.55
5	0	81	596.55	640.71
6	0	81	640.71	665.98
7	0	81	665.98	666.46
8	0	81	666.46	673.42
9	0	81	674.86	701.84
10	0	81	702.05	725.64

Figure 7 MS/MS Search running

Exercise 6. Autovalidate the hiqhest-quality results

After you have completed a database search, you validate the good results. Validation means that you accept that the matches are correct.

The Spectrum Mill workbench provides a means to segregate search results that contain a valid interpretation of an MS/MS spectrum from those that do not. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

The Spectrum Mill workbench provides two ways to validate results. One way uses the Autovalidation page, and is totally automated. You use this method only to validate the highest-scoring results — those that do not require manual review. The other method of validating uses the Protein/Peptide Summary page for manual review and validation. This exercise describes autovalidation.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 8 on page 31.

2 Check that your **Data Directory** is set to **heatshock**.	• If you have just performed the data extraction and MS/MS Search, your **heatshock** data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Keep the default scoring presets, as shown in Figure 8. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein. • Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20.

Steps	Detailed Instructions	Comments
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Keep the default scoring presets, as shown in Figure 9 on page 32. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by peptide. Even if it finds only a single peptide for a protein, it validates the corresponding search results provided that the peptide score is high enough. • Using the default scoring, individual peptides must have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein.

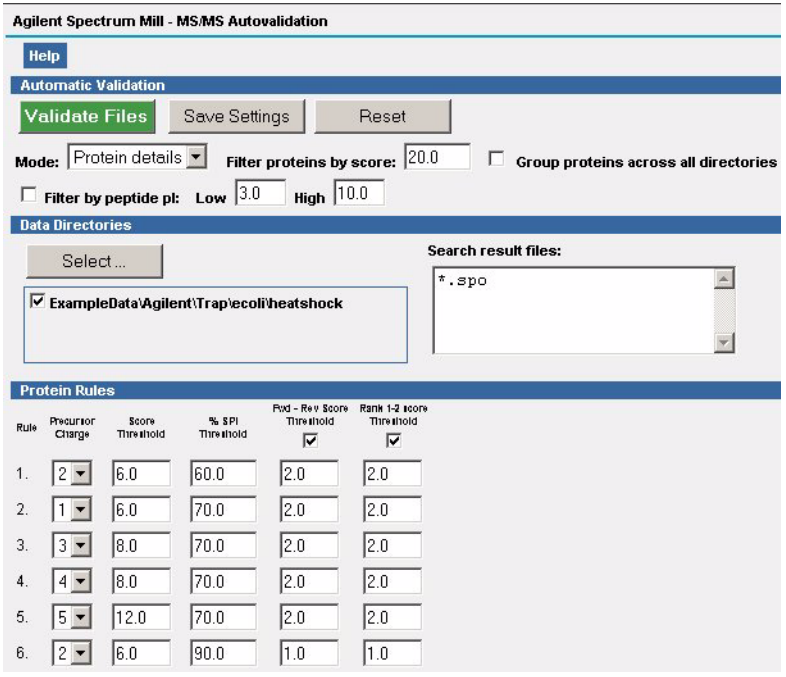


Figure 8 Autovalidation form, Protein details mode

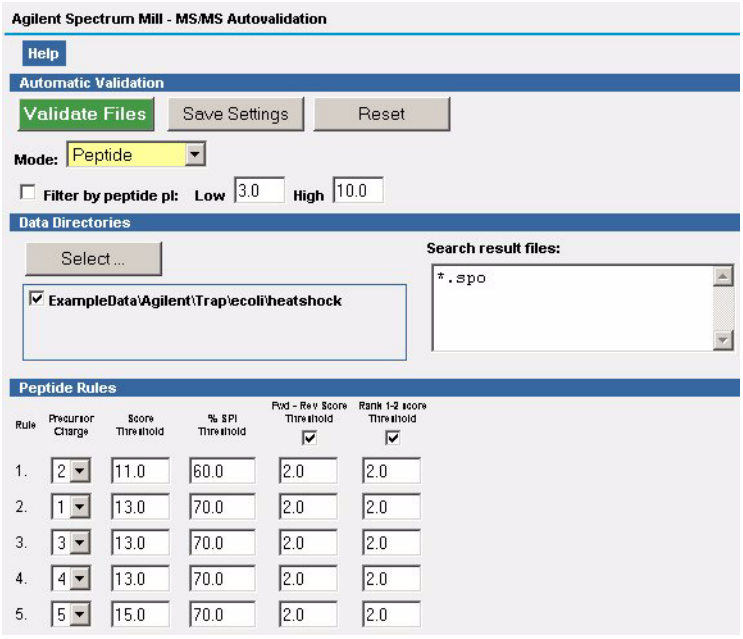


Figure 9 Autovalidation form, Peptide mode

Exercise 7. (Optional) Examine the validated results

If this is the first time you have performed an autovalidation, you may want to review the validated data to convince yourself that autovalidation really works.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page, if you are not already there.	• From the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other Spectrum Mill pages.
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just performed autovalidation, your **heatshock** data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Set the **Validation preset** to **none**.		

Steps	Detailed Instructions	Comments
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	By setting to zero, you ensure that all valid data is selected, regardless of score.
8 Set the **Review Fields**.	**a** Mark the check box for **Rank 1-2 score**. **b** For others, keep the default settings. **c** Check to see that your form looks like that in Figure 10. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		

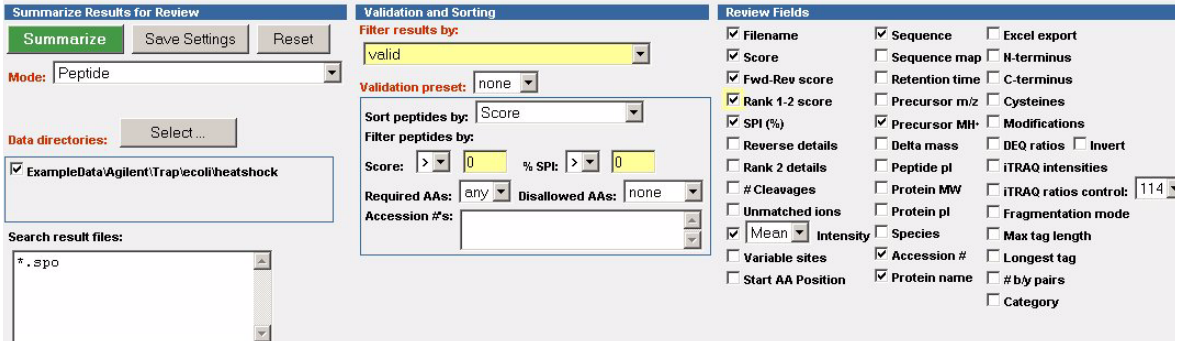


Figure 10 Settings to check autovalidation results

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 380 **Hits:** 380

#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	Spectrum Intensity	Sequence	MH+ Matched (Da)	Accession #	Protein Name
1	HEC_0040.1927.1927.0	3	25.53	25.53	25.53	96.3	1.20e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
2	HEC_0040.1867.1879.0	3	25.25	25.25	25.25	91.0	3.51e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
3	HEC_0010.1397.1405.2	2	24.54	21.35	24.54	97.4	9.49e+006	(K)DVTTGDTLCDPDAPIILER(M)	2101.012	15803853	elongation factor EF-2
4	HEC_0020.1546.1555.2	2	24.27	19.09	16.40	98.0	6.15e+007	(R)TILVPIDISDSELTQR(V)	1799.975	15801753	hypothetical protein Z2335
5	HEC_0250.1283.1286.0	3	24.24	24.24	20.28	98.0	3.90e+006	(K)VLESAIANAEHNDGADIDDLK(V)	2210.057	15803842	50S ribosomal protein L22
6	HEC_0010.1669.1675.2	2	23.83	23.83	18.04	93.9	4.80e+006	(K)DIQLATPPQVGAPATEYAALAELK(A)	2467.308	16130805	glycine decarboxylase, PLP-dependent, subu
7	HEC_0100.1323.1332.0	3	23.47	23.47	23.47	97.7	6.53e+006	(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	2894.410	16129055	malonyl-CoA-[acyl-carrier-protein] transacyla
8	HEC_0020.1307.1319.0	2	23.37	23.37	23.37	93.6	2.75e+007	(K)DTTTIIDGVGEEAAIQGR(V)	1845.919	15804735	chaperonin GroEL
9	HEC_0020.1259.1259.0	2	22.84	22.84	22.84	94.7	2.70e+007	(K)DTTTIIDGVGEEAAIQGR(V)	1845.919	15804735	chaperonin GroEL
10	HEC_0020.1363.1367.2	2	22.24	18.80	22.24	92.1	3.35e+007	(K)DVTTGDTLCDPDAPIILER(M)	2101.012	15803853	elongation factor EF-2
11	HEC_0010.1547.1562.2	2	22.18	22.18	22.18	89.8	8.29e+006	(K)DNPPQDLIDLNPNQSVPTLVDR(E)	2460.237	15803763	stringent starvation protein A
12	HEC_0020.0995.1009.2	2	22.12	15.76	16.33	98.6	1.14e+008	(K)AIAQVGTISANSDETVGK(L)	1760.902	18028150	GroEL
13	HEC_0010.1372.1384.2	2	22.06	22.06	22.06	92.7	1.71e+007	(K)DVTTGDTLCDPDAPIILER(M)	2101.012	15803853	elongation factor EF-2
14	HEC_0060.1397.1397.2	2	21.86	21.86	21.86	91.9	9.33e+006	(R)AAPSYEELSNSQELLETGIK(V)	2179.077	15804332	ATP synthase subunit B
15	HEC_0020.0886.0898.0	2	21.78	14.59	13.57	96.1	1.71e+008	(K)TGEVPADVAAQAR(Q)	1284.654	15804834	hypothetical protein Z5854
16	HEC_0020.0836.0840.2	2	21.72	15.99	15.72	96.1	6.93e+007	(K)ASSGLNEDEIQK(M)	1290.617	15799694	molecular chaperone DnaK
17	HEC_0080.1311.1311.0	3	21.69	16.12	16.89	95.2	7.73e+006	(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	2894.410	16129055	malonyl-CoA-[acyl-carrier-protein] transacyla
18	HEC_0020.1390.1399.2	2	21.61	18.35	21.61	94.7	4.49e+007	(R)GDLGVEIGDPELVGIQK(A)	1738.922	16129807	pyruvate kinase II
19	HEC_0020.1679.1682.0	2	21.60	21.60	21.60	98.5	2.34e+007	(K)GLTDAAQQVVAAVEGK(-)	1556.828	15800432	succinyl-CoA synthetase subunit beta
20	HEC_0020.0789.0792.2	2	21.45	12.55	21.45	97.3	8.13e+006	(R)NVVAGDPSPDGQGR(L)	1368.650	48783	ProRS

Figure 11 Summary of valid results from autovalidation (partial list)

Steps	Detailed Instructions	Comments
10 Examine the summary report.	**a** Check that your results are similar to those in Figure 11 on page 35. Since the public databases are constantly updated, your results may vary slightly. **b** Become familiar with the following headers: • **#** - table entry number • **Filename** - the name of the spectral file • **z** - peptide charge • **Score** - peptide database search score • **Fwd-Rev Score** - difference between scores for top hits from forward and reversed database searches • **Rank1-Rank2 Score** - difference between rank 1 and rank 2 database search scores • **SPI (%)** - % scored peak intensity, which is the percentage of the ion current in the extracted MS/MS spectrum that is explained by the top database hit • **Spectrum Intensity** - intensity calculated from extracted ion chromatogram of precursor ion • **Sequence** - sequence from top database hit • **MH^+ Matched (Da)** - measured precursor ion MH^+ • **Accession #** - database accession number for the top hit • **Protein Name** - name of the top database hit	File names are in the format: Data_File_Name.aaaa.bbbb.c, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)
11 To examine these results in more detail, go to "Exercise 10. Examine results with the Spectrum Viewer" on page 44.		

Exercise 8. Display medium-quality results in preparation for manual validation

Now that you have automatically validated the best results with autovalidation, you prepare to validate medium-quality results through a process of manual inspection and validation. Manual validation helps you to maximize the number of proteins you identify in your samples, or to increase the sequence coverage of the identified proteins, but it does require that you devote some time. When you process your own samples, you may skip this step if automatic validation provides sufficient protein identifications for the purposes of your experiment.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button near the top of the form to display the settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other pages.
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just performed autovalidation or results review, your **heatshock** data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Steps	Detailed Instructions	Comments
4 For **Filter results by**, reset to the default setting of **sequence-not-validated**.		• **Filter results by** selects from your data only the results that match your setting. • In this case, it selects only the results that have not yet been validated during autovalidation.
5 Set the **Validation preset** to **reset**.		• You always want to set the **Validation preset** to the setting most likely to be accurate for the majority of the data. This minimizes manual changes later. • Since you will set score thresholds to select the medium-quality results, the **reset** setting is best for this round of processing. That way you manually change the **Validation category** setting to **valid** (**V**) for those matches you accept as being correct.
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** For **Score**, keep the default of **> 6**, to capture all results that may possibly be good for ion trap data. **b** For **% SPI**, keep the default of **> 60**. Lower values suggest poor results.	• These parameters select from the data set only the members that exhibit the specified score criteria. • These settings select ion trap results of medium quality or better. • If these were Agilent Q-TOF results, you would use a **Score** setting of **> 5** and a **% SPI** setting of **> 60**. • Remember that the **Filter results by** setting prevents display of the highest-quality results (the ones that have already been validated).

Steps	Detailed Instructions	Comments
8 Set the **Review Fields**.	**a** Mark the check box for **Rank 1-2 score**. **b** Mark the check boxes for **# Cleavages** and **Unmatched ions**. **c** For others, keep the default settings. **d** Check to see that your form looks like that in Figure 12. Note that settings that have been changed from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. Differences of at least 1 and preferably 2 can indicate a good match, especially for the lower-scoring hits. • The **# Cleavages** is the number of cleavages of the amino acid backbone that are represented in the spectrum for the assigned sequence. For CID spectra, these cleavages include the b- and y-fragments, and for ETD spectra, they include the c- and z-fragments. Other ions, such as neutral losses from b and y ions, are not included. • When you use **Dynamic peak thresholding** in the search it is helpful to mark the check box for **Unmatched ions**. These results tell you both the number of unmatched ions and the number of ions that were included in the search score that you see in the report. If the latter number is less than 15, check that the peptide is either short or in low abundance. If not, be suspicious of the results. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		

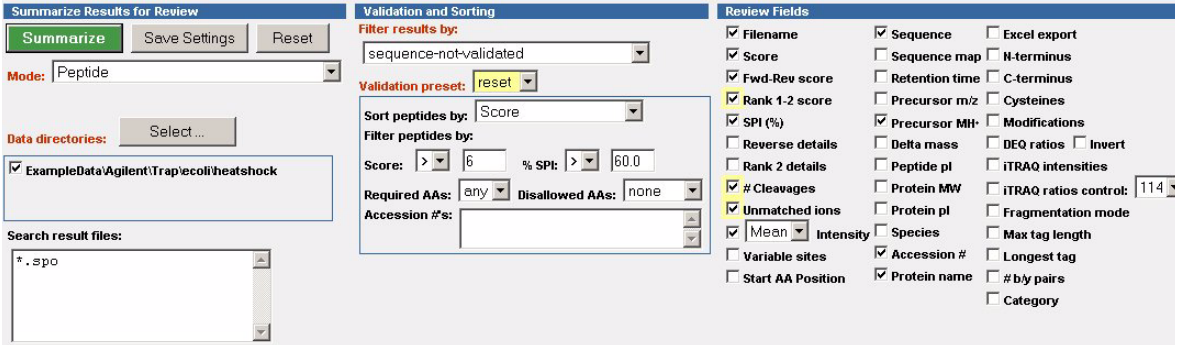
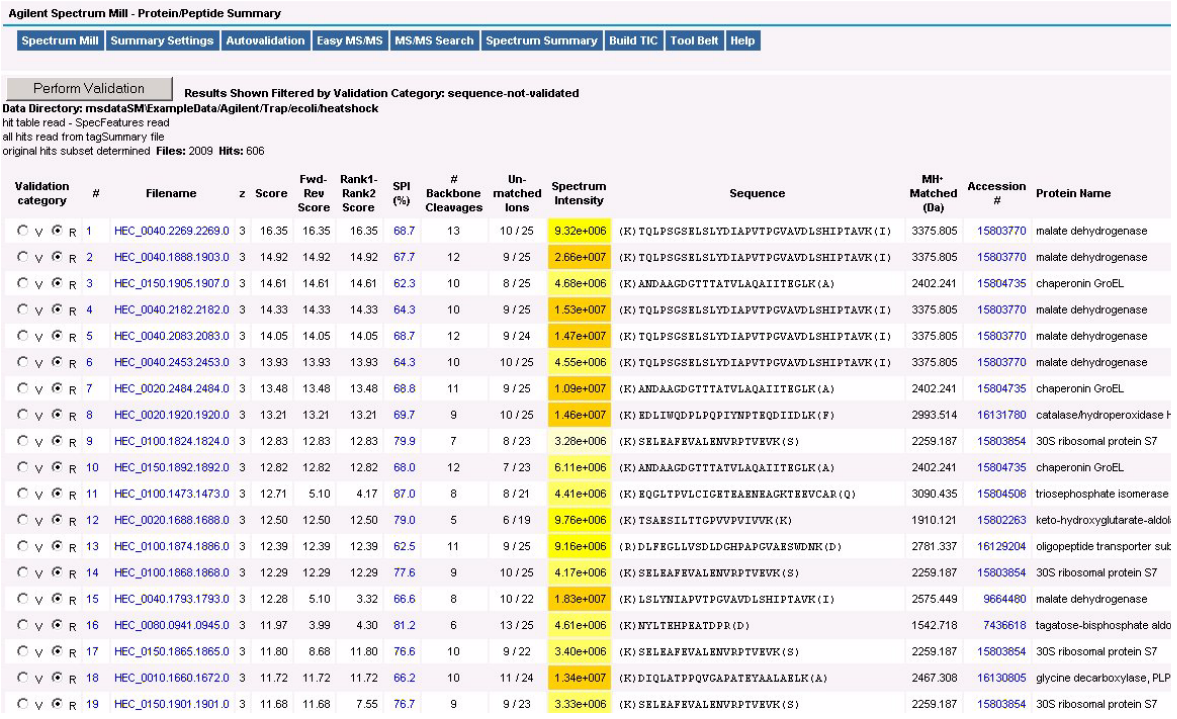


Figure 12 Protein/Peptide Summary settings to display medium-quality results

Steps	Detailed Instructions	Comments
10 Examine the summary report.	**a** Check to see that your results are similar to those in Figure 13. Note that since the public databases are constantly updated, your results may vary slightly.	File names are in the format: Data_File_Name.aaaa.bbbb.c, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)
11 To examine these results in more detail, go to "Exercise 9. Manually review and validate results" on page 42.		
12 (Optional) Vary input parameters and investigate the results.	**a** To change parameters and create a new summary table, click the **Summary Settings** button near the top of the form to start over. **b** Before proceeding to the next exercise, be sure to go back to the settings you had in step 8 of this exercise.	



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Perform Validation **Results Shown Filtered by Validation Category: sequence-not-validated**

Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
all hits read from tagSummary file
original hits subset determined **Files:** 2009 **Hits:** 606

Validation category	#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	# Backbone Cleavages	Un-matched Ions	Spectrum Intensity	Sequence	MH+ Matched (Da)	Accession #	Protein Name
○ V ◉ R	1	HEC_0040.2269.2269.0	3	16.35	16.35	16.35	68.7	13	10 / 25	9.32e+006	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
○ V ◉ R	2	HEC_0040.1888.1903.0	3	14.92	14.92	14.92	67.7	12	9 / 25	2.66e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
○ V ◉ R	3	HEC_0150.1905.1907.0	3	14.61	14.61	14.61	62.3	10	8 / 25	4.68e+006	(K)ANDAAGDGTTTATVLAQAIITEGLK(A)	2402.241	15804735	chaperonin GroEL
○ V ◉ R	4	HEC_0040.2182.2182.0	3	14.33	14.33	14.33	64.3	10	9 / 25	1.53e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
○ V ◉ R	5	HEC_0040.2083.2083.0	3	14.05	14.05	14.05	68.7	12	9 / 24	1.47e+007	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
○ V ◉ R	6	HEC_0040.2453.2453.0	3	13.93	13.93	13.93	64.3	10	10 / 25	4.55e+006	(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)	3375.805	15803770	malate dehydrogenase
○ V ◉ R	7	HEC_0020.2484.2484.0	3	13.48	13.48	13.48	68.8	11	9 / 25	1.09e+007	(K)ANDAAGDGTTTATVLAQAIITEGLK(A)	2402.241	15804735	chaperonin GroEL
○ V ◉ R	8	HEC_0020.1920.1920.0	3	13.21	13.21	13.21	69.7	9	10 / 25	1.46e+007	(K)EDLIWQDPLPQPIYNPTEQDIIDLK(F)	2993.514	16131780	catalase/hydroperoxidase H
○ V ◉ R	9	HEC_0100.1824.1824.0	3	12.83	12.83	12.83	79.9	7	8 / 23	3.28e+006	(K)SELEAFEVALENVRPTVEVK(S)	2259.187	15803854	30S ribosomal protein S7
○ V ◉ R	10	HEC_0150.1892.1892.0	3	12.82	12.82	12.82	68.0	12	7 / 23	6.11e+006	(K)ANDAAGDGTTTATVLAQAIITEGLK(A)	2402.241	15804735	chaperonin GroEL
○ V ◉ R	11	HEC_0100.1473.1473.0	3	12.71	5.10	4.17	87.0	8	8 / 21	4.41e+006	(K)EQGLTPVLCIGETEAENEAGKTEEVCAR(Q)	3090.435	15804508	triosephosphate isomerase
○ V ◉ R	12	HEC_0020.1688.1688.0	3	12.50	12.50	12.50	79.0	5	6 / 19	9.76e+006	(K)TSAESILTTGPVVPVIVVK(K)	1910.121	15802263	keto-hydroxyglutarate-aldol
○ V ◉ R	13	HEC_0100.1874.1886.0	3	12.39	12.39	12.39	62.5	11	9 / 25	9.16e+006	(R)DLFEGLLVSDLDGHPAPGVAESWDNK(D)	2781.337	16129204	oligopeptide transporter sub
○ V ◉ R	14	HEC_0100.1868.1868.0	3	12.29	12.29	12.29	77.6	9	10 / 25	4.17e+006	(K)SELEAFEVALENVRPTVEVK(S)	2259.187	15803854	30S ribosomal protein S7
○ V ◉ R	15	HEC_0040.1793.1793.0	3	12.28	5.10	3.32	66.6	8	10 / 22	1.83e+007	(K)LSLYNIAPVTPGVAVDLSHIPTAVK(I)	2575.449	9664480	malate dehydrogenase
○ V ◉ R	16	HEC_0080.0941.0945.0	3	11.97	3.99	4.30	81.2	6	13 / 25	4.61e+006	(K)NYLTEHPEATDPR(D)	1542.718	7436618	tagatose-bisphosphate aldo
○ V ◉ R	17	HEC_0150.1865.1865.0	3	11.80	8.68	11.80	76.6	10	9 / 22	3.40e+006	(K)SELEAFEVALENVRPTVEVK(S)	2259.187	15803854	30S ribosomal protein S7
○ V ◉ R	18	HEC_0010.1660.1672.0	3	11.72	11.72	11.72	66.2	10	11 / 24	1.34e+007	(K)DIQLATPPQVGAPATEYAALAELK(A)	2467.308	16130805	glycine decarboxylase, PLP
○ V ◉ R	19	HEC_0150.1901.1901.0	3	11.68	11.68	7.55	76.7	9	9 / 23	3.33e+006	(K)SELEAFEVALENVRPTVEVK(S)	2259.187	15803854	30S ribosomal protein S7

Figure 13 Summary of medium-quality results (partial list)

Exercise 9. Manually review and validate results

After you display medium-quality results, you manually review them. In this exercise, you examine the MS/MS Search results to determine how well they fit the experimental spectra. As you examine the results, you designate good results as **valid**.

Steps	Detailed Instructions	Comments
1 Display the first spectrum.	**a** In the summary report, click the number **1** under the **#** header. **b** Check the bottom of the browser window, where you see the first spectrum displayed in the Spectrum Viewer. **c** Check that your results are similar to those in Figure 14 on page 43. Since the public databases are constantly updated, your results may vary. **d** Now turn your attention again to the summary report. **e** Click the first number under the **SPI** header. **f** Check that you see the same spectrum displayed in the Spectrum Viewer as you did in substep **b**, only this time the Spectrum Viewer appears in a new window.	The Spectrum Viewer shows your extracted MS/MS spectrum. This spectrum is annotated with theoretical fragments from the peptide identified as the best database match.
2 Use the Spectrum Viewer to examine details of the MS/MS search interpretation and its correlation with the spectrum.	**a** Learn to use the Spectrum Viewer. Follow the steps in "Exercise 10. Examine results with the Spectrum Viewer" on page 44. **b** Use the Spectrum Viewer to evaluate the search results.	In the spectrum, unmatched ions are shown in black, while matched ions are displayed in red and blue.

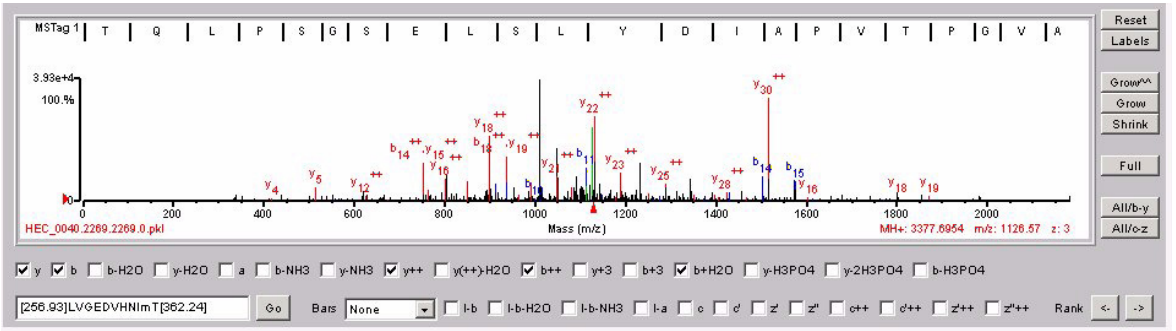

MSTag 1 T Q L P S G S E L S L Y D I A P V T P G V A
Reset
Labels
Grow^^
Grow
Shrink
Full
All/b-y
All/c-z
3.93e+4
100.%
0
0 200 400 600 800 1000 1200 1400 1600 1800 2000
Mass (m/z)
HEC_0040.2269.2269.0.pkl
MH+: 3377.6954 m/z: 1126.57 z: 3
y b b-H2O y-H2O a b-NH3 y-NH3 y++ y(++)-H2O b++ y+3 b+3 b+H2O y-H3PO4 y-2H3PO4 b-H3PO4
[256.93]LVGEDVHNImT[362.24]
Go
Bars None
I-b I-b-H2O I-b-NH3 I-a c c' z' z'' c++ c'++ z'++ z''++
Rank
<-
->

Figure 14 Spectrum Viewer

Steps	Detailed Instructions	Comments
3 (Optional) Display the MS/MS Search results and evaluate details.	• Follow instructions in "Exercise 13. (Optional) Examine details of search results" on page 55.	
4 Decide whether to mark the database search results for the first spectrum as **valid**.	**a** For purposes of this exercise, you decide that the protein sequence is a good match for the spectrum. **b** Under the **Validation category** heading, click the **V** to set the validation preset to **valid**.	The match you see in this exercise may actually be mediocre since autovalidation finds most of the good matches.
5 Examine additional spectra.	**a** Repeat step 2 (and optionally step 3) for as many spectra as you wish to examine. **b** Do not change any more of the validation presets under the **Validation category** heading. If you do, your results will not match in later exercises.	• When you process your own data, you will likely review all the spectra. • When you process your own data, you will change the **Validation category** presets from **reset** to **valid** (**R** to **V**) as you encounter good results.
6 Click the **Perform Validation** button.	• Check that you see a validation summary that lists all the hits and spectra that have been validated thus far (from both autovalidation and manual validation).	This makes your assessments a permanent part of the validation record.

Exercise 10. Examine results with the Spectrum Viewer

The Spectrum Viewer is a visual tool to evaluate MS/MS search results (Protein/Peptide Summary), to assess MS/MS spectral quality (Spectrum Summary), and to evaluate *de novo* sequencing results (Sherenga *de novo* Summary). This exercise focuses on use of the Spectrum Viewer for evaluation of MS/MS search results. In this capacity, the Spectrum Viewer shows your extracted spectrum annotated with theoretical fragments from the top database search result. This viewer has flexible scaling and labeling options to allow you to better visualize your results.

Steps	Detailed Instructions	Comments
1 If you have not already done so, display a Protein/Peptide Summary report.	• Follow steps in "Exercise 8. Display medium-quality results in preparation for manual validation" on page 37.	
2 If you have not already done so, display the first spectrum.	**a** In the summary report, click the number **1** under the **#** header. **b** Check the bottom of the browser window, where you see the first spectrum displayed in the Spectrum Viewer. **c** Check that your results are similar to those in Figure 14 on page 43. Since the public databases are constantly updated, your results may vary.	

Steps	Detailed Instructions	Comments
3 Use the **Labels** and **All/b-y** buttons to change the spectrum annotations.	**a** Click several times on the **Labels** button on the right side of the Spectrum Viewer. **b** Check that the peak labeling changes as you click. **c** Be sure the labels are visible before proceeding. **d** Click twice on the **All/b-y** button on the right side of the Spectrum Viewer. **e** Check that as you click, the marks in the check boxes immediately below the spectrum change, as do the spectrum labels.	• In the spectrum, unmatched ions are shown in black, while matched ions are displayed in red and blue. • The **Labels** button toggles among different peak labeling options. These are: ◦ The default of interpreted peaks (b, y, etc.) ◦ Interpreted peaks plus mass labels of *all* peaks ◦ Interpreted peaks plus mass labels of interpreted peaks ◦ No labels • The **All/b-y** button toggles marking of check boxes in the first row under the spectrum and labels peaks accordingly. The toggle either marks all the check boxes or resets to defaults.
4 Use the **Grow**, **Grow^^**, **Shrink**, and **Reset** buttons to expand and shrink the spectrum in the vertical (y) direction.	**a** Click twice on the **Grow** button on the right side of the Spectrum Viewer. **b** Check that the spectrum expands vertically. **c** Click twice on the **Shrink** button on the right side of the Spectrum Viewer. **d** Check that the spectrum returns to original size. **e** Click twice again on the **Grow** button. **f** Click once on the **Reset** button on the right side of the Spectrum Viewer. **g** Click once on the **Grow^^** button. **h** Click once on the **Reset** button. **i** Check that the spectrum returns to original size.	

Steps	Detailed Instructions	Comments
5 Use the mouse and the **Full** and **Reset** buttons to expand and shrink the spectrum in the horizontal (x) direction.	**a** Use the cursor to expand a portion of the x-axis of the spectrum. Move your mouse over the spectrum. When a crosshair is displayed, select the portion of the spectrum you wish to expand. **b** Double-click the spectrum or click the **Reset** button to return to the original display.	• The **Full** button displays the spectrum with the full x-axis values. • If you see a black arrow at the lower right-hand corner of the spectrum, this may mean that some of the spectrum is not displayed because there were no significant peaks in that region. Click the **Full** button to display the full x-axis range of the spectrum.
6 Select **Bars** options to locate all potential b- and y-ions.	**a** Find the **Bars** options at the bottom of the Spectrum Viewer. **b** Select **y** and observe the gray lines that indicate locations of all possible y-ions from the database match. **c** Select **b** and observe the gray lines that indicate locations of all possible b-ions from the database match. Notice that the amino acid sequence reverses order when you switch from **y** to **b**. **d** Select **None** to return to the default setting.	By choosing **y** or **b** for **Bars**, you quickly see how well the sequence from the summary table matches the spectrum.
7 Type in a sequence and check for alignment with the spectrum.	**a** Find the sequence box in the lower left-hand corner of the Spectrum Viewer. **b** Type in your own sequence. **c** Click the **Go** button. **d** Select the **b** and **y Bars** options and observe whether your trial sequence is consistent with the mass information in the spectrum. **e** Click the left **Rank** arrow to go back to the database sequence. 	• This feature is useful when you know part of a sequence, or you want to test a possible sequence. • The sequence in the box is initially set to a default. • Mass gaps shown in brackets indicate segments of the sequence that are unknown. Perhaps there are portions of your MS/MS spectrum where fragmentation is insufficient to provide an amino acid sequence. • You can enter mass gaps in the middle of the sequence as well as at the ends.

Exercise 11. Summarize valid results

Now you summarize all the results you have validated thus far. These include results validated both automatically and manually.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button near the top of the page to display settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just performed results review, your **heatshock** data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Set display **Mode** to **Protein Summary**.		• When you process your own samples, you may choose another display mode, depending on whether you want to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set the **Filter results by** to **valid**.		• **Filter results by** selects from your data only the results that match your setting. • In this case, it selects only the valid results.

Steps	Detailed Instructions	Comments
5 For **Sort proteins by**, keep the default of **Score**.		• The **Sort** parameters determine how the data are sorted in the results display. • When you process your own data, select the setting that is most helpful to you.
6 Set the **score** and **% SPI** filters to **> 0**.	**a** Set **Filter by protein score** to **> 0**. **b** For **Filter peptides by**, set both **Score** and **% SPI** to **> 0**.	When you set these filters to **> 0**, you are sure to display all the valid results.
7 Set the **Review Fields**.	**a** Mark the **Protein MW** check box. **b** Keep the other default settings. **c** Check to see that your form looks like that in Figure 15. Note that settings that are changed from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
8 Click the **Summarize** button.		

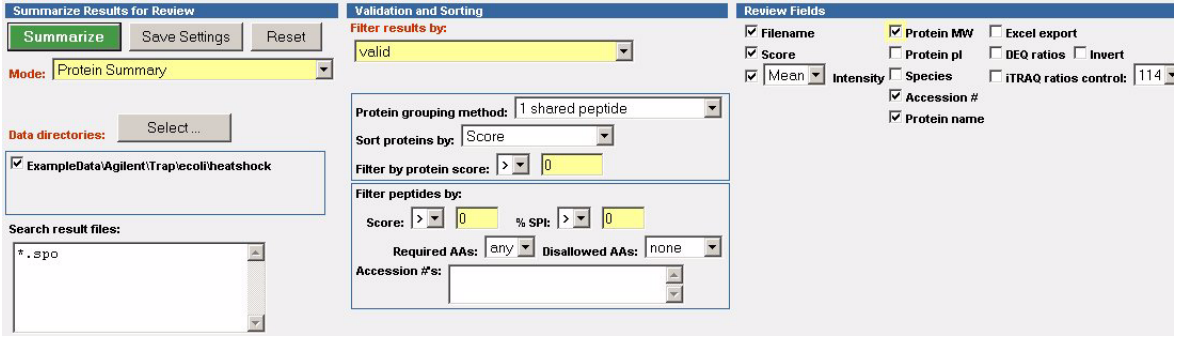


Figure 15 Protein/Peptide Summary settings to display a final results summary

Steps	Detailed Instructions	Comments
9 Examine the overall summary report.	• Check that your results are similar to those in Figure 16 on page 50.	• Since the public databases are constantly updated, your results may vary, • Your results also vary depending on which peptides you manually validated.
10 Note the colored cells in the summary report.		• The color code in the **Distinct Peptides** column indicates the relative number of peptides detected for each protein. Darker colors represent larger numbers of peptides. • The color code in the **Mean Peptide Spectral Intensity** column indicates relative protein concentrations. Darker colors represent higher concentrations. The number in each cell is an average of the intensities for all the peptides assigned to that protein. The peptide intensities are calculated from extracted ion chromatograms from the precursor ions. • For more accurate quantitation, under **Review Fields**, for **Intensity**, select **Total** rather than **Mean**. • The intensity results are sufficient for studies where you are interested in differences of two-fold or more.
11 (Optional) Vary the summary settings (particularly the **Mode** setting) and investigate the results.		See the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 381 **Hits:** 2654
beginning to assemble proteins .... proteins assembled 0.413511 sec
proteins filtered by unique peptides 0.148062 sec
proteins filtered by score
calculated protein coverage maps 0.451565 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.202965 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 67 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Protein MW (Da)	Database Accession #	Protein Name
1	101	6	114.14	18	1.33e+007	57357.2	45686198	GroEL
2	11	6	92.25	17	1.68e+007	69115.3	15799694	molecular chaperone DnaK
3	18	4	77.77	9	2.98e+007	77581.7	15803853	elongation factor EF-2
4	6	3	48.79	15	4.12e+007	31090.6	15803363	5-keto-4-deoxyuronate isomerase
5	11	3	48.47	9	8.59e+006	48384.3	15800756	seryl-tRNA synthetase
6	13	3	46.72	21	3.12e+007	20761.5	15800320	alkyl hydroperoxide reductase, C22 subunit; detoxification of hyd
7	8	3	46.28	8	1.91e+007	56976.9	15804515	glycerol kinase
8	28	2	44.22	13	2.50e+007	35057.7	26249817	malate dehydrogenase
9	21	2	36.55	6	3.15e+007	44850.6	26249935	elongation factor Tu
10	5	2	35.37	10	1.77e+007	42384.1	15803459	fructose-bisphosphate aldolase
11	5	2	34.91	25	1.48e+007	11299.9	26248730	Hypothetical protein yfcZ
12	5	2	34.76	8	1.63e+007	47325.2	15804766	adenylosuccinate synthetase
13	2	2	34.60	5	9.55e+006	63701.9	48783	ProRS
14	8	2	32.82	15	8.76e+006	13099.5	15803825	30S ribosomal protein S13
15	5	2	31.71	8	2.09e+007	43298.0	26249339	phosphoglycerate kinase
16	3	2	31.62	4	2.25e+007	66096.4	16128108	pyruvate dehydrogenase, dihydrolipoyltransacetylase compone
17	4	2	29.60	2	2.58e+007	151886.6	26250758	DNA-directed RNA polymerase beta subunit
18	2	2	28.02	10	2.07e+007	35742.8	15802706	galactose-binding transport protein; receptor for galactose taxis

Figure 16 Final results summary (partial listing), in Protein Summary mode

Exercise 12. Compare results from two samples

The example data includes two *E. coli* samples — a "heatshock" (stressed) sample and a "wild" (control) sample. The goal of this exercise is to compare the proteins in the two samples, to see which are up-regulated and which are down-regulated in the stressed cells. To compare the two, you first need to process the "wild" sample in the same way you processed the "heatshock" sample. To save time, the exercise limits you to a few key processing steps for the "wild" sample. If time permits, you may apply all the previous exercises to the "wild" sample.

Steps	Detailed Instructions	Comments
1 Extract the **ecoli\wild** data.	• Repeat "Exercise 3. Extract the spectra" on page 16, except when you **Select** the **Data Directory**, select **wild** rather than **heatshock**.	
2 Search the **ecoli\wild** data.	• Repeat "Exercise 5. Search the database" on page 24, except when you **Select** the **Data Directory**, select **wild** rather than **heatshock**.	
3 Autovalidate the **ecoli\wild** data.	• Repeat "Exercise 6. Autovalidate the hiqhest-quality results" on page 30, except when you **Select** the **Data Directory**, select **wild** rather than **heatshock**.	

Steps	Detailed Instructions	Comments
4 Summarize and compare the **ecoli\heatshock** and the **ecoli\wild** data.	**a** Repeat "Exercise 11. Summarize valid results" on page 47, with the following changes: • When you **Select** the **Data Directory**, select *both* **heatshock** and **wild**. (Control-click the second directory. Do not use shift-click.) • For **Mode**, select **Protein-Protein Comparison Columns**. • For **Group results by**, choose **Directory**. **b** Check to see that the completed form looks like that in Figure 17. **c** Click the **Summarize** button.	The **Group results by** parameter controls how the sample results are displayed. • If you select **File**, each sample fraction is displayed individually. • If you select **Directory**, all fractions present in a data directory are consolidated into a single sample.

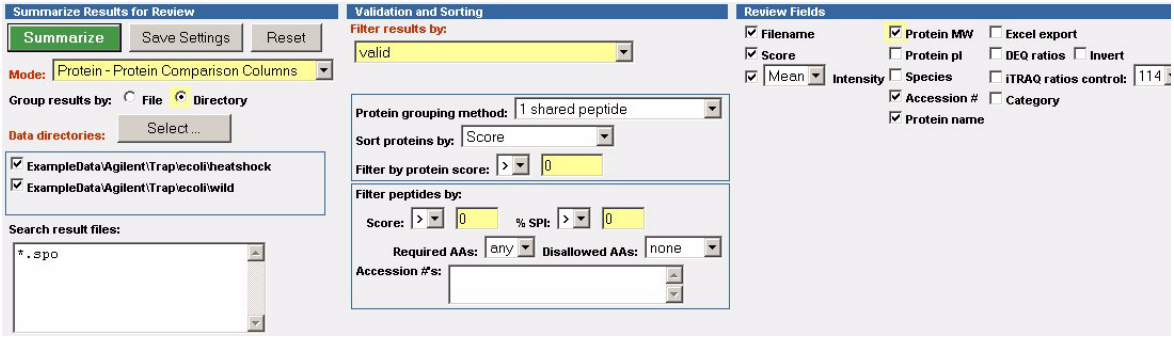
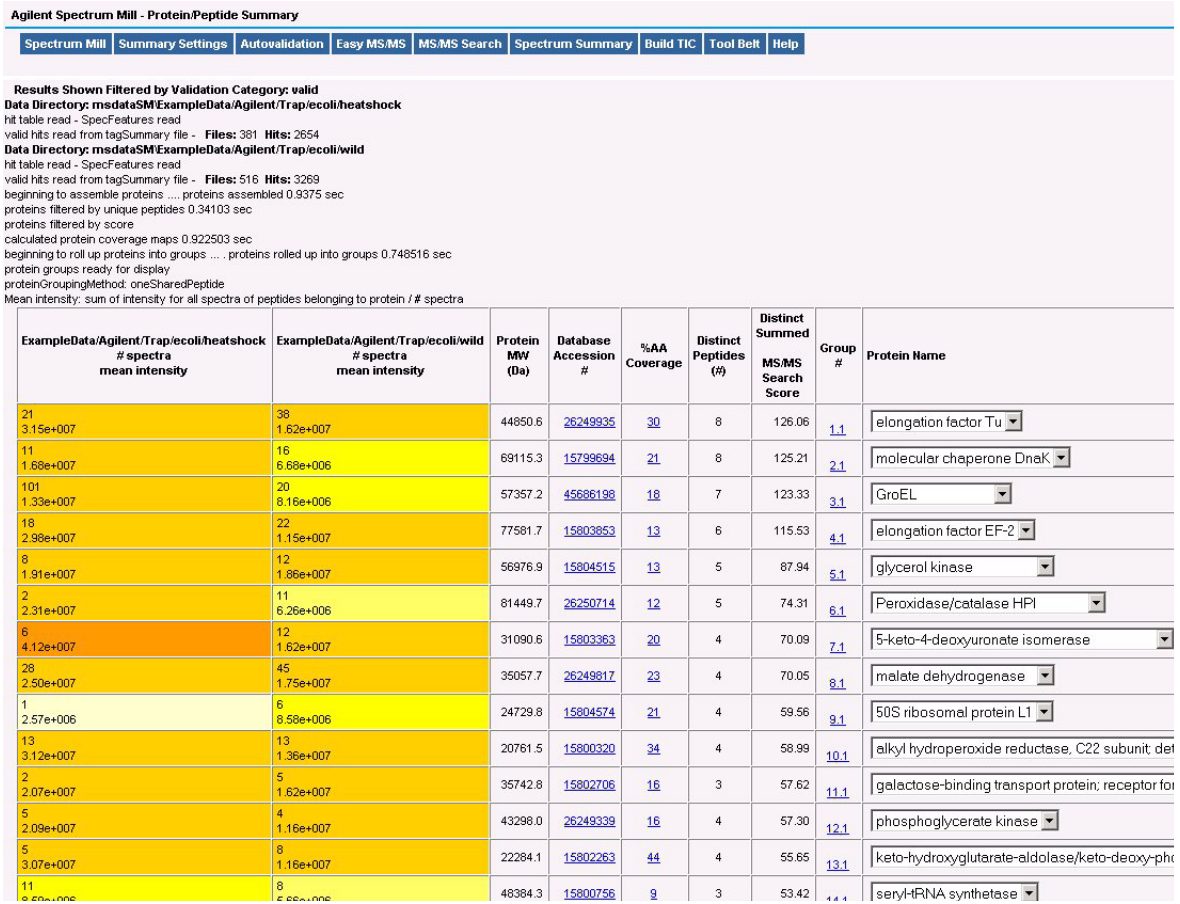


Figure 17 Protein/Peptide Summary settings to compare two samples

Steps	Detailed Instructions	Comments

5 Examine the overall summary report.	• Check that your results are similar to those in Figure 18 on page 54.
	  ◦ Since the public databases are constantly updated, your results may vary,
	  ◦ Your results also vary depending on which peptides you manually validated.
		Note how the colored cells in the summary report make it easy to visualize sample differences.
		• The color-code in the left columns represents the relative protein concentrations. Darker colors represent higher concentrations.
		• The top number in each cell is the number of peptides detected for each protein, while the bottom number is the mean peptide intensity, an average of the intensities for all peptides detected for that protein. These intensities are calculated from extracted ion chromatograms from the precursor ions.
		• These intensity results are sufficient for studies where you are interested in differences of two-fold or more.

6 (Optional) Vary the summary settings (particularly the **Mode** setting) and investigate the results.		See the MS/MS Data Review and Validation chapter of the *Spectrum Mill MS Proteomics Workbench Application Guide*.



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 381 **Hits:** 2654
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/wild
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 516 **Hits:** 3269
beginning to assemble proteins ... proteins assembled 0.9375 sec
proteins filtered by unique peptides 0.34103 sec
proteins filtered by score
calculated protein coverage maps 0.922503 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.748516 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide
Mean intensity: sum of intensity for all spectra of peptides belonging to protein / # spectra

ExampleData/Agilent/Trap/ecoli/heatshock # spectra mean intensity	ExampleData/Agilent/Trap/ecoli/wild # spectra mean intensity	Protein MW (Da)	Database Accession #	%AA Coverage	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	Group #	Protein Name
21 3.15e+007	38 1.62e+007	44850.6	26249935	30	8	126.06	1.1	elongation factor Tu
11 1.68e+007	16 6.68e+006	69115.3	15799694	21	8	125.21	2.1	molecular chaperone DnaK
101 1.33e+007	20 8.16e+006	57357.2	45686198	18	7	123.33	3.1	GroEL
18 2.98e+007	22 1.15e+007	77581.7	15803853	13	6	115.53	4.1	elongation factor EF-2
8 1.91e+007	12 1.86e+007	56976.9	15804515	13	5	87.94	5.1	glycerol kinase
2 2.31e+007	11 6.26e+006	81449.7	26250714	12	5	74.31	6.1	Peroxidase/catalase HPI
6 4.12e+007	12 1.62e+007	31090.6	15803363	20	4	70.09	7.1	5-keto-4-deoxyuronate isomerase
28 2.50e+007	45 1.75e+007	35057.7	26249817	23	4	70.05	8.1	malate dehydrogenase
1 2.57e+006	6 8.58e+006	24729.8	15804574	21	4	59.56	9.1	50S ribosomal protein L1
13 3.12e+007	13 1.36e+007	20761.5	15800320	34	4	58.99	10.1	alkyl hydroperoxide reductase, C22 subunit; det
2 2.07e+007	5 1.62e+007	35742.8	15802706	16	3	57.62	11.1	galactose-binding transport protein; receptor fo
5 2.09e+007	4 1.16e+007	43298.0	26249339	16	4	57.30	12.1	phosphoglycerate kinase
5 3.07e+007	8 1.16e+007	22284.1	15802263	44	4	55.65	13.1	keto-hydroxyglutarate-aldolase/keto-deoxy-pho
11	8	48384.3	15800756	9	3	53.42		seryl-tRNA synthetase

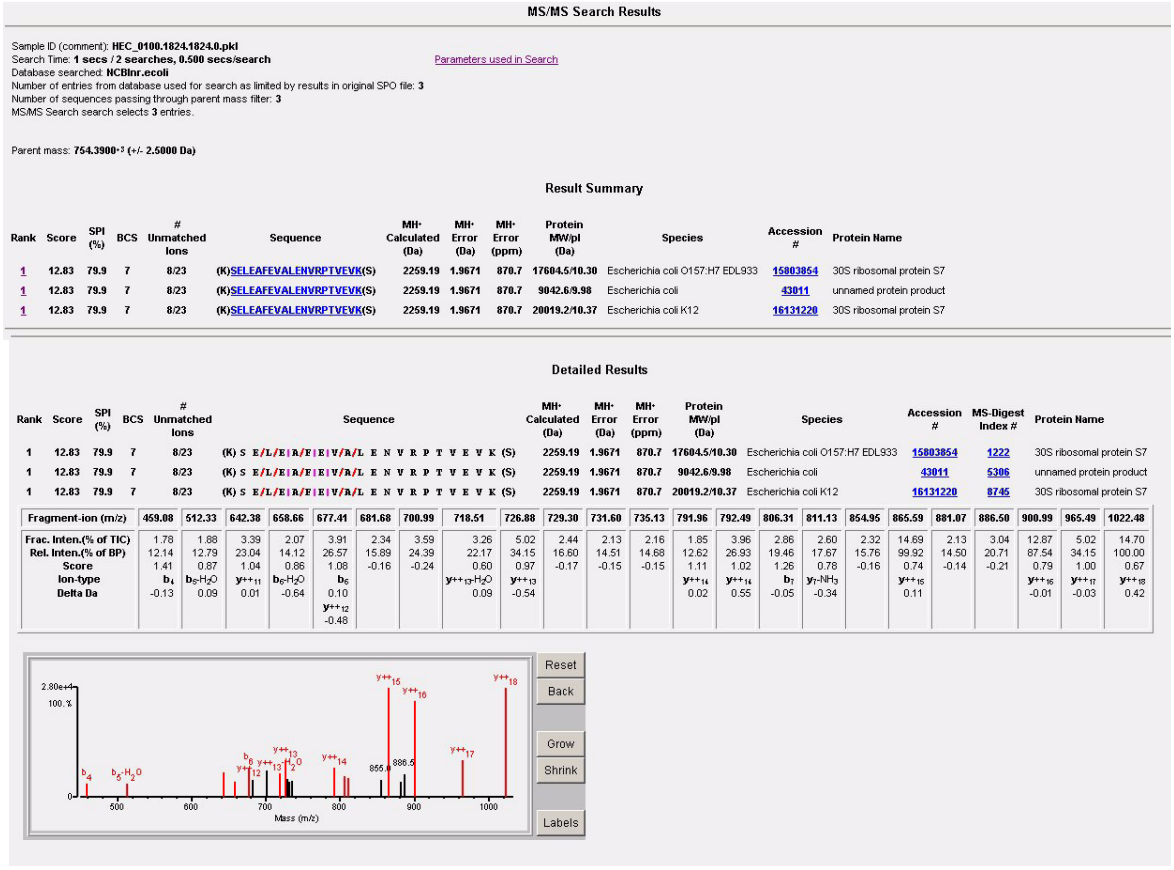
Figure 18 Comparison of proteins from two *E. coli* cell states (partial listing)

Exercise 13. (Optional) Examine details of search results

While the Spectrum Viewer is your primary tool to determine how well your database hit matches your spectrum, you may occasionally want to examine search results in more detail. In this exercise, you become familiar with the additional database search details that are available.

Steps	Detailed Instructions	Comments
1 If you have not already done so, display a Protein/Peptide Summary report.	• Follow steps in "Exercise 8. Display medium-quality results in preparation for manual validation" on page 37.	
2 Display the MS/MS Search results.	**a** In the summary report, click the first entry under the **Filename** header. **b** Check the bottom of the browser window, where you see the MS/MS Search results for the first spectrum. See Figure 19 on page 57 as an example.	• At this point, you may see only the spectrum. • The example in Figure 19 on page 57 is different than the example in Figure 14 on page 43. This is because the hit list for the latter was too long to display on a single page of this manual.
3 Expand the section of the page that shows the MS/MS Search results.	**a** To adjust the size of this section, rest your mouse over the split line. **b** When the pointer becomes a double-sided arrow, drag the arrow to move the split line.	

Steps	Detailed Instructions	Comments
4 Examine the MS/MS Search hit list.	**a** Scroll up to see the hit list at the top of the MS/MS Search results section (top of Figure 19 on page 57.) **b** Note the color-coded lines in the sequence and read the comment to the right in this table. **c** Note that since the public databases are constantly updated, your results may vary.	• The color-coding of the amino acid sequence of the matched peptide from the database search shows: ◦ Red forward-slashes for locations of y-ions ◦ Blue backslashes for locations of b-ions ◦ Magenta pipes (vertical lines) for locations of both b- and y-ions • This allows you to assess b- and y-ion coverage without inspecting the spectrum, and for variable modifications results aids in identifying the site of a post-translational modification (PTM) by highlighting existence of surrounding b/y ions.



MS/MS Search Results

Sample ID (comment): **HEC_0100.1824.1824.0.pkl**
Search Time: **1 secs / 2 searches, 0.500 secs/search** Parameters used in Search
Database searched: **NCBInr.ecoli**
Number of entries from database used for search as limited by results in original SPO file: **3**
Number of sequences passing through parent mass filter: **3**
MS/MS Search search selects **3** entries.

Parent mass: **754.3900+3 (+/- 2.5000 Da)**

Result Summary

Rank	Score	SPI (%)	BCS	# Unmatched Ions	Sequence	MH+ Calculated (Da)	MH+ Error (Da)	MH+ Error (ppm)	Protein MW/pI (Da)	Species	Accession #	Protein Name
1	12.83	79.9	7	8/23	(K)SELEAFEVALENVRPTVEVK(S)	2259.19	1.9671	870.7	17604.5/10.30	Escherichia coli O157:H7 EDL933	15803854	30S ribosomal protein S7
1	12.83	79.9	7	8/23	(K)SELEAFEVALENVRPTVEVK(S)	2259.19	1.9671	870.7	9042.6/9.98	Escherichia coli	43011	unnamed protein product
1	12.83	79.9	7	8/23	(K)SELEAFEVALENVRPTVEVK(S)	2259.19	1.9671	870.7	20019.2/10.37	Escherichia coli K12	16131220	30S ribosomal protein S7

Detailed Results

Rank	Score	SPI (%)	BCS	# Unmatched Ions	Sequence	MH+ Calculated (Da)	MH+ Error (Da)	MH+ Error (ppm)	Protein MW/pI (Da)	Species	Accession #	MS-Digest Index #	Protein Name
1	12.83	79.9	7	8/23	(K) S E/L/E/A/F/E/V/A/L E N V R P T V E V K (S)	2259.19	1.9671	870.7	17604.5/10.30	Escherichia coli O157:H7 EDL933	15803854	1222	30S ribosomal protein S7
1	12.83	79.9	7	8/23	(K) S E/L/E/A/F/E/V/A/L E N V R P T V E V K (S)	2259.19	1.9671	870.7	9042.6/9.98	Escherichia coli	43011	5306	unnamed protein product
1	12.83	79.9	7	8/23	(K) S E/L/E/A/F/E/V/A/L E N V R P T V E V K (S)	2259.19	1.9671	870.7	20019.2/10.37	Escherichia coli K12	16131220	8745	30S ribosomal protein S7

Fragment-ion (m/z)	459.08	512.33	642.38	658.66	677.41	681.68	700.99	718.51	726.88	729.30	731.60	735.13	791.96	792.49	806.31	811.13	854.95	865.59	881.07	886.50	900.99	965.49	1022.48
Frac. Inten.(% of TIC)	1.78	1.88	3.39	2.07	3.91	2.34	3.59	3.26	5.02	2.44	2.13	2.16	1.85	3.96	2.86	2.60	2.32	14.69	2.13	3.04	12.87	5.02	14.70
Rel. Inten.(% of BP)	12.14	12.79	23.04	14.12	26.57	15.89	24.39	22.17	34.15	16.60	14.51	14.68	12.62	26.93	19.46	17.67	15.76	99.92	14.50	20.71	87.54	34.15	100.00
Score	1.41	0.87	1.04	0.86	1.08	-0.16	-0.24	0.60	0.97	-0.17	-0.15	-0.15	1.11	1.02	1.26	0.78	-0.16	0.74	-0.14	-0.21	0.79	1.00	0.67
Ion-type	b_4	b_5-H_2O	y^{++}_{11}	b_6-H_2O	b_6			y^{++}_{13}-H_2O	y^{++}_{13}				y^{++}_{14}	y^{++}_{14}	b_7	y_7-NH_3		y^{++}_{15}			y^{++}_{16}	y^{++}_{17}	y^{++}_{18}
Delta Da	-0.13	0.09	0.01	-0.64	0.10			0.09	-0.54				0.02	0.55	-0.05	-0.34		0.11			-0.01	-0.03	0.42
					y^{++}_{12} -0.48																		

Figure 19 MS/MS Search results

Steps	Detailed Instructions	Comments
5 View details of the search results.	**a** Scroll below the hit list to see details of the search results (Figure 19). **b** Inspect the table for the rank-one hit. The table shows in detail how the ions were assigned. Note the following: • **Fragment-ion (m/z)**: ions from the experimental spectrum • **Frac. Inten. (% of TIC)**: fractional intensities for each of the fragment ions. The final % SPI for the database match is the sum of the fractional intensities for the assigned ions. • **Rel. Inten. (% of BP)**: relative intensities for each of the fragment ions, measured as a percentage of the base peak • **Score**: scores for each of the fragment ions. The final score for the database match is the sum of the scores for each of the ions. (The unassigned ions subtract from the score.) • **Ion-type**: ion type for theoretical fragment from database match • **Delta Da**: difference between theoretical and experimental MH^+ **c** Inspect the spectrum just below the table to see visually how well the interpretation matches the spectrum.	

Steps	Detailed Instructions	Comments
6 (Optional) Scroll back up to the top of the MS/MS Search Result Summary table to discover how all the links work.	• To jump to a particular search result, click a number under the **Rank** heading in the **Result Summary**. • To view an MS Product listing that shows all theoretical fragment ions, click an amino acid sequence in the **Result Summary**. • To view information in the database you searched, click an accession number in the **Result Summary**. • To view MS Digest results that show peptide masses from a theoretical digestion, click an MS Digest index number in the **Detailed Results**.	• When you check links, click your browser's **back** arrow to return to the previous display. • When you process your data, you only click the links you feel are necessary to understand the quality of the results. These steps are presented here only so you know what information is available to help you with your decision.
7 Navigate back to the main Protein/Peptide Summary report.		You may need to readjust the page partition so you can see the top section again.
8 (Optional) Click the remaining link here to see what it does.	• To link to the database you searched, click a link under the **Accession #** header.	If you wish to enable a link from the amino acid sequence to a BLAST database, see the system administration chapter in the *Application Guide*.

2
Optional Processing for MS/MS Data

Exercise 1. Create a list of identified proteins for variable modifications search 62
Exercise 2. Search in variable modifications mode 64
Exercise 3. Review and validate variable modifications search results 69
Exercise 4. Check for remaining high-quality spectra 72
Exercise 5. Perform Sherenga de novo sequencing 77
Exercise 6. Review results from Sherenga de novo sequencing 80

This chapter provides additional familiarization exercises to process an example set of ion trap MS/MS data files with the Agilent Spectrum Mill MS Proteomics Workbench. These exercises correspond with the optional processing in the second half of Chapter 1 of the companion *Spectrum Mill MS Proteomics Workbench Application Guide*. How much of the optional processing you do depends on how important it is to extract every last bit of information from your sample analyses. Since the exact steps you take are sample- and study-dependent, you are encouraged to vary the processing to fit your samples and information needs. When you are satisfied with the amount of information you have gleaned from your samples, you generate a final results summary, as described in Chapter 1.

As you continue to process your data, Spectrum Mill retains a cumulative summary of all the results you have validated. At the end, your final results summary includes everything you have validated through multiple interpretation/review/validation steps.

While the exercises in this chapter use Agilent ion trap data files, the general workflow applies broadly for MS/MS data files.

Exercise 1. Create a list of identified proteins for variable modifications search

The MS/MS Search software provides identity, variable modifications, and homology search modes. In identity mode, you search for exact matches in the selected database. In variable modifications mode, you search for post-translational and chemical modifications that occur at a fraction of the available sites. In homology modes, you search for single amino acid substitutions relative to the database (as well as any variable modifications you have selected). In homology modes, you can also search for an unknown or unexpected modification. For details, see Chapter 5, “Searching for Unknown Modifications”.

As you attempt to extract additional information from your sample, one option is to search a second database in identity mode, as described in “Exercise 5. Search the database” on page 24. For example, if you first searched against the **SwissProt** database, try the larger **NCBInr** database. If you could not find hits in the standard protein databases, try searching **dbEST**.

For this series of exercises, you perform a variable modifications search. You can perform such a search against a full database, a species-specific subset database, or a list of proteins that you have already identified in your sample. In this series of exercises, you search against the latter. This saves time because the modified peptides you find in variable modifications mode are often derived from the proteins you already identified in identity mode. The variable modifications search is much faster when you search a smaller list.

While not described in detail here, you may also conduct a no enzyme search against the list of proteins that you have already identified in your sample. Such a search identifies any partial tryptic or non-tryptic peptides.

This exercise shows how to create a file of previously validated hits, which is a small database of proteins that you have already identified in your sample. In this exercise, you create the file using the Tool Belt page. Alternatively, you can create the file automatically from the MS/MS Search page. When you mark the check box to **Search previous hits** in MS/MS Search, if you have not created a saved results file, then the search page automatically creates one for you. (On the form, be sure to select the database that you searched previously.)

Steps	Detailed Instructions	Comments
1 Navigate to the Tool Belt page.	• From the Spectrum Mill home page, click the link to the Tool Belt page.	There are other ways to navigate to the Tool Belt page. Check the top of your current page for a **Tool Belt** button.
2 Click the **Create saved results file** option near the top of the form.		• You will create a file of previously validated hits from the search results from "Exercise 5. Search the database" on page 24. • You will later search this protein list in variable modifications mode.
3 Fill in the form as shown in Figure 20.	**a** Select the data directory that you previously searched (**heatshock**). **b** Select the database you previously searched (**NCBInr.ecoli**).	
4 Click the **Create File** button.		

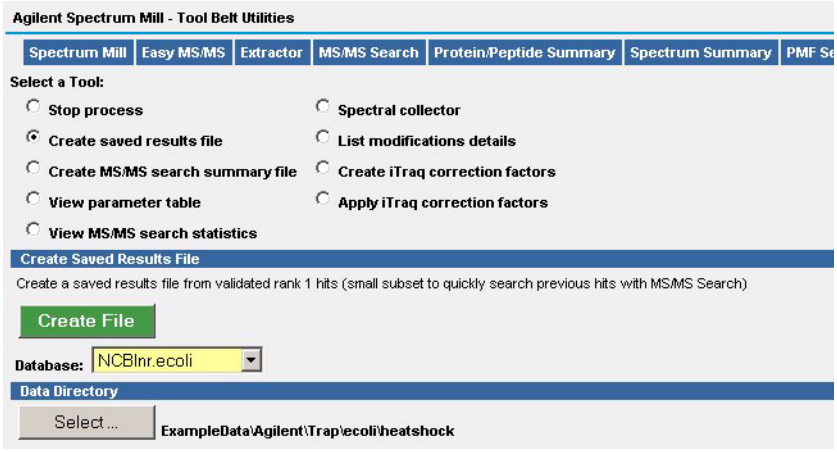


Figure 20 Create file of previously validated proteins

Exercise 2. Search in variable modifications mode

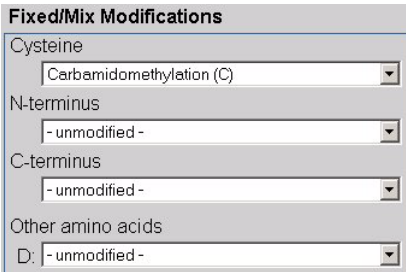
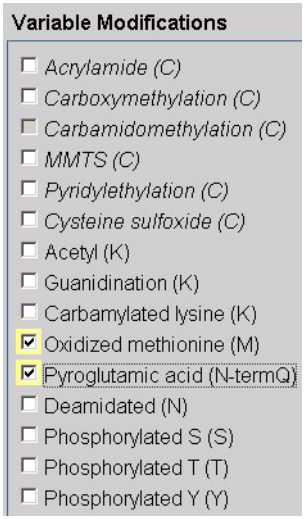
In this exercise, you search in variable modifications mode against database hits that you validated previously. When you process your own samples, you could instead search in variable modifications mode against a full or species subset database. If you search a species subset database, you may want to start with a well-defined species (e.g., human), and then do a second variable modifications search on a broader range of species (e.g., mammal). The broader the database search, the longer the search time. A search of previous hits as in this exercise is very efficient.

This exercise describes a single search in variable modifications mode. When you process your own samples, you may conduct multiple searches, with autovalidation and manual validation between each set of conditions:

1. Combine the following common modifications into a single search: oxidized methionine (methionine sulfoxide) and pyroglutamic acid.
2. If you suspect phosphorylation, combine the following modifications into a single search: phosphorylated S, phosphorylated T, and phosphorylated Y.
3. Search for any other modifications you suspect for your sample.

Steps	Detailed Instructions	Comments
1 Navigate to the MS/MS Search page.	• From the Tool Belt page or the Spectrum Mill home page, click the link to MS/MS Search.	There are a number of ways to navigate to the MS/MS Search page.
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just been working with this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Set **Search Parameters**.	**a** For **Database**, select **NCBInr.ecoli**, as before. **b** Mark the check box next to **Search previous hits.**	This searches the file of previous hits you created for this data directory. (See "Exercise 1. Create a list of identified proteins for variable modifications search" on page 62.)

Steps	Detailed Instructions	Comments

4 Choose the appropriate modifications.	**a** Click the **Choose...** button near the middle of the form. **b** Choose **Fixed/Mix Modifications**. Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Choose **Variable Modifications**. Select **Oxidized methionine** and **Pyroglutamic acid**.  **d** Click **OK.** The names of the modifications appear as in Figure 21 on page 68.	• The **fixed** modifications are assumed to apply universally and are searched in a single search cycle.
		• The **mix** modifications trigger cyclic MS/MS searches, where a different form of the modification is searched in each cycle.
		• The **variable** modifications allow for both modified and unmodified forms within a peptide. All variable modifications are searched in each cycle.
		• Here we choose two variable modifications that occur commonly:
		  ◦ Oxidation of methionine (M to m)
		  ◦ Peptide N-terminal glutamine to pyroglutamic acid (Q to q)
		• Note that when you select a variable modification, the **Search mode** changes automatically to **Variable modifications**.
		• For more information about choosing modifications, see the online help.
		• Your system administrator can configure custom modifications.

Steps	Detailed Instructions	Comments
5 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Proton mobility scoring**. **c** Mark the check box for **Dynamic peak thresholding**. **d** Verify that the **Search mode** is set to **Variable modifications**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Protein mobility scoring** is a scoring enhancement for ion trap data that applies a bonus or penalty based on the expected relative intensities for MS/MS fragments from a given peptide sequence. When you process your own ion trap data, it is best to use this scoring, except with modifications that significantly alter the fragmentation patterns of peptides. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.
6 Set the other parameters as shown in Figure 21 on page 68.	**a** Keep the defaults whenever appropriate. Changes from defaults are highlighted in yellow. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	• You do not mark the check box to **Remove all prior MS/MS results** because you want to combine the results of this search with the results of the previous one. You mark this check box only if you feel you have searched with incorrect settings or if for any other reason you want to start over.

Steps	Detailed Instructions	Comments
7 Start the search.	**a** Click the **Start Search** button. **b** View search progress in the area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • Search time varies depending on the size of the database searched. • For the same data set and database, variable modifications and homology searches take longer than identity searches because there are more permutations to consider. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

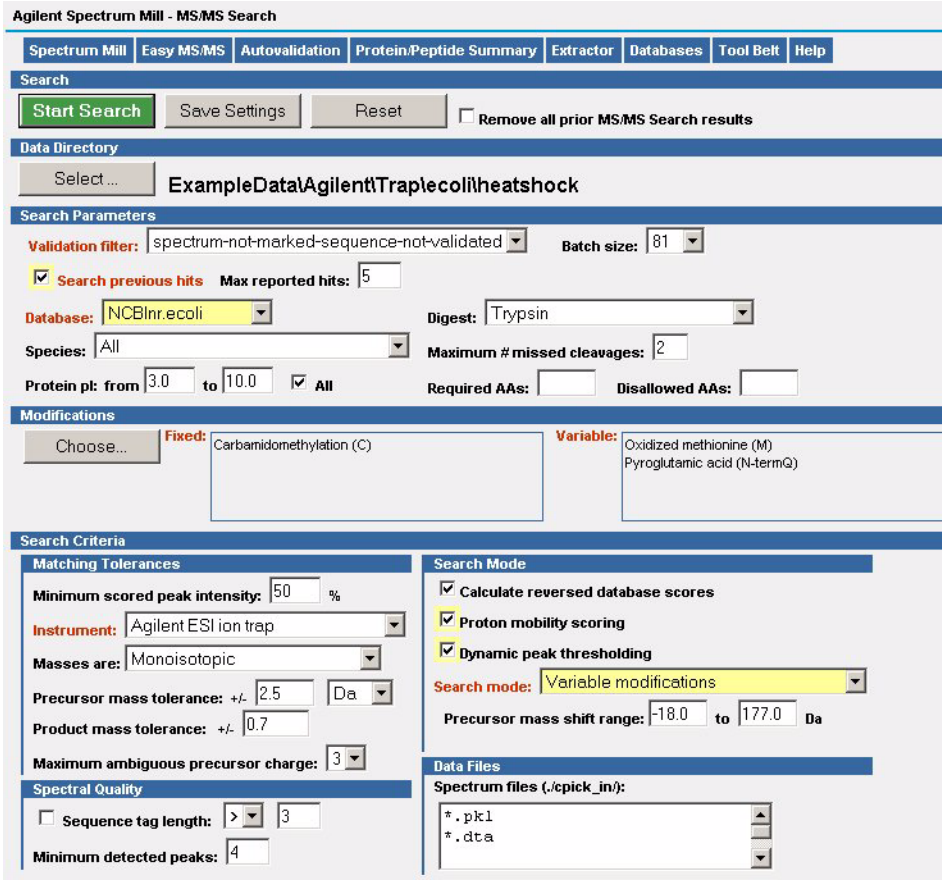


Figure 21 MS/MS Search settings for search of previous hits for variable modifications

Exercise 3. Review and validate variable modifications search results

You review and validate variable modifications search results in much the same way as identity mode search results.

Steps	Detailed Instructions	Comments
1 Perform autovalidation on the variable modifications search results.	• Follow the procedure in "Exercise 6. Autovalidate the hiqhest-quality results" on page 30.	• You may use autovalidation with searches that are limited to a few variable modifications. • Autovalidation is not recommended for the homology modes because of the increased possibility of false positives when so many permutations are considered.
2 (Optional) Examine the validated variable modifications search results.	• Follow the procedure in "Exercise 7. (Optional) Examine the validated results" on page 33, with the following exceptions: ◦ To see the variable modifications on the Protein/Peptide Summary page, under **Review Fields**, mark the **Variable sites**, **Delta mass**, and **Modifications** check boxes, as shown in Figure 22 on page 70. ◦ Check that the data display includes the additional headings **Variable Sites**, **Modifications**, **MH^+ Mass Shift (Da)**, and **MH^+ Error (ppm)**, as shown in Figure 23 on page 71. ◦ The columns for **Variable Sites** and **MH^+ Mass Shift (Da)** summarize the variable modifications.	• This summary includes results validated from both identity mode and variable modifications searches. • You may need to scroll down the results table to see the variable modifications, which are indicated as lower-case letters in the column for **Variable Sites**. • Tip: If you want to see the variable modifications grouped at the top and/or bottom of the results table, for **Sort peptides by**, select **Delta mass shift**. • If the variable modifications occur on the N-terminus, C-terminus, or cysteines, additionally mark the appropriate check box(es) in the last column of the **Review Fields**.

Steps	Detailed Instructions	Comments
3 Display medium-quality variable modifications search results in preparation for manual validation.	• Follow the procedure in "Exercise 8. Display medium-quality results in preparation for manual validation" on page 37, with the following exceptions: ◦ Under **Review Fields**, mark the **Variable sites**, **Delta mass** and **Modifications** check boxes. These allow you see the variable modifications on the Protein/Peptide Summary page. ◦ To see the variable modifications grouped at the top and/or bottom of the results table, for **Sort peptides by**, select **Delta mass shift**.	
4 Manually review and validate variable modifications search results.	• Follow the procedure in "Exercise 9. Manually review and validate results" on page 42, but review only the results that show variable modifications.	You may need to scroll down the results table to see the variable modifications, which are indicated as lower-case letters in the column for **Variable Sites**.
5 Create a final results summary.	• Follow the procedure in "Exercise 11. Summarize valid results" on page 47.	This summary includes results that you validated from both identity mode and variable modifications searches.

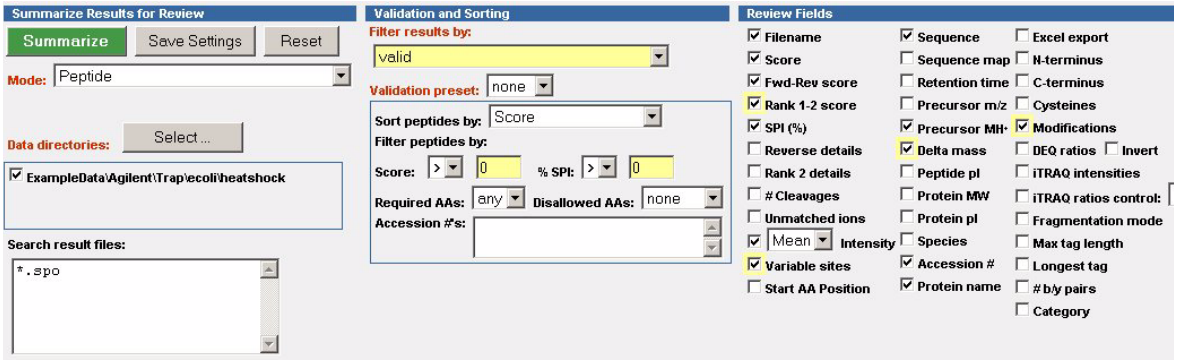
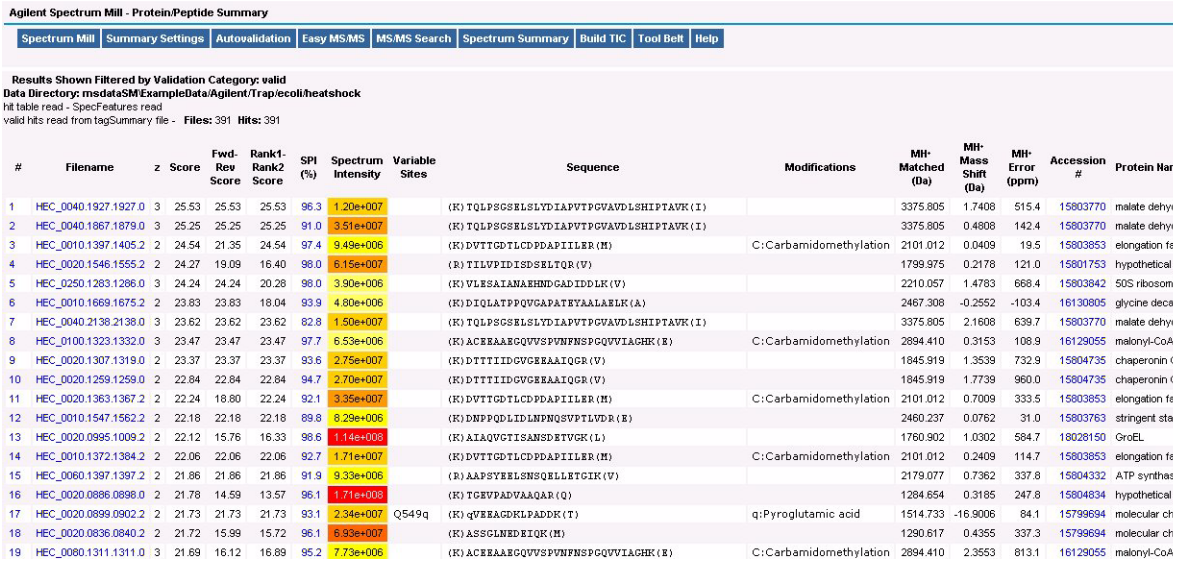


Figure 22 Settings to display additional review fields for variable modifications search results



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ecoli/heatshock
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 391 **Hits:** 391

#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	Spectrum Intensity	Variable Sites	Sequence	Modifications	MH+ Matched (Da)	MH+ Mass Shift (Da)	MH+ Error (ppm)	Accession #	Protein Nar
1	HEC_0040.1927.1927.0	3	25.53	25.53	25.53	96.3	1.20e+007		(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)		3375.805	1.7408	515.4	15803770	malate dehy
2	HEC_0040.1867.1879.0	3	25.25	25.25	25.25	91.0	3.51e+007		(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)		3375.805	0.4808	142.4	15803770	malate dehy
3	HEC_0010.1397.1405.2	2	24.54	21.35	24.54	97.4	9.49e+006		(K)DVTTGDTLCDPDAPIILER(M)	C:Carbamidomethylation	2101.012	0.0409	19.5	15803853	elongation fa
4	HEC_0020.1546.1555.2	2	24.27	19.09	16.40	98.0	6.15e+007		(R)TILVPIDISDSELTQR(V)		1799.975	0.2178	121.0	15801753	hypothetical
5	HEC_0250.1283.1286.0	3	24.24	24.24	20.28	98.0	3.90e+006		(K)VLESAIANAEHNDGADIDDLK(V)		2210.057	1.4783	668.4	15803842	50S ribosom
6	HEC_0010.1669.1675.2	2	23.83	23.83	18.04	93.9	4.80e+006		(K)DIQLATPPQVGAPATEYAALAELK(A)		2467.308	-0.2552	-103.4	16130805	glycine deca
7	HEC_0040.2138.2138.0	3	23.62	23.62	23.62	82.8	1.50e+007		(K)TQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVK(I)		3375.805	2.1608	639.7	15803770	malate dehy
8	HEC_0100.1323.1332.0	3	23.47	23.47	23.47	97.7	6.53e+006		(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	C:Carbamidomethylation	2894.410	0.3153	108.9	16129055	malonyl-CoA
9	HEC_0020.1307.1319.0	2	23.37	23.37	23.37	93.6	2.75e+007		(K)DTTTIIDGVGEEAAIQGR(V)		1845.919	1.3539	732.9	15804735	chaperonin (
10	HEC_0020.1259.1259.0	2	22.84	22.84	22.84	94.7	2.70e+007		(K)DTTTIIDGVGEEAAIQGR(V)		1845.919	1.7739	960.0	15804735	chaperonin (
11	HEC_0020.1363.1367.2	2	22.24	18.80	22.24	92.1	3.35e+007		(K)DVTTGDTLCDPDAPIILER(M)	C:Carbamidomethylation	2101.012	0.7009	333.5	15803853	elongation fa
12	HEC_0010.1547.1562.2	2	22.18	22.18	22.18	89.8	8.29e+006		(K)DNPPQDLIDLNPNQSVPTLVDR(E)		2460.237	0.0762	31.0	15803763	stringent sta
13	HEC_0020.0995.1009.2	2	22.12	15.76	16.33	98.6	1.14e+008		(K)AIAQVGTISANSDETVGK(L)		1760.902	1.0302	584.7	18028150	GroEL
14	HEC_0010.1372.1384.2	2	22.06	22.06	22.06	92.7	1.71e+007		(K)DVTTGDTLCDPDAPIILER(M)	C:Carbamidomethylation	2101.012	0.2409	114.7	15803853	elongation fa
15	HEC_0060.1397.1397.2	2	21.86	21.86	21.86	91.9	9.33e+006		(R)AAPSYEELSNSQELLETGIK(V)		2179.077	0.7362	337.8	15804332	ATP synthas
16	HEC_0020.0886.0898.0	2	21.78	14.59	13.57	96.1	1.71e+008		(K)TGEVPADVAAQAR(Q)		1284.654	0.3185	247.8	15804834	hypothetical
17	HEC_0020.0899.0902.2	2	21.73	21.73	21.73	93.1	2.34e+007	Q549q	(K)qVEEAGDKLPADDK(T)	q:Pyroglutamic acid	1514.733	-16.9006	84.1	15799694	molecular ch
18	HEC_0020.0836.0840.2	2	21.72	15.99	15.72	96.1	6.93e+007		(K)ASSGLNEDEIQK(M)		1290.617	0.4355	337.3	15799694	molecular ch
19	HEC_0080.1311.1311.0	3	21.69	16.12	16.89	95.2	7.73e+006		(K)ACEEAAEGQVVSPVNFNSPGQVVIAGHK(E)	C:Carbamidomethylation	2894.410	2.3553	813.1	16129055	malonyl-CoA

Figure 23 Valid search results from both identity mode and variable modifications searches. The results from the variable modifications search are indicated by entries in the **Variable Sites** column.

Exercise 4. Check for remaining high-quality spectra

In this exercise, you learn to use the Spectrum Summary page to view spectra that do not yet have a valid interpretation, and to decide whether they are good spectra worthy of further interpretation.

One of the most useful measures of spectral quality is the maximum sequence tag length. The maximum sequence tag length represents the longest sequence of amino acids that can be located in the spectrum. The **Maximum tag length** filter retains spectra that are likely to be peptide spectra and that have a fragmentation pattern that lends itself to interpretation. The Spectrum Summary results show spectra annotated with the amino acid sequences that correspond with the maximum sequence tag lengths. These are not intended as *de novo* interpretations, but rather show how much sequence information *could* be contained in the spectra. This exercise illustrates use of this filter to locate remaining good spectra.

Steps	Detailed Instructions	Comments
1 Navigate to the Spectrum Summary page.	• Navigate to this page from the Spectrum Mill home page, the Protein/Peptide Summary page or the Tool Belt page.	
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just been working with this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Leave the **Spectrum validation filter** set to the default of **spectrum-not-marked-sequence-not-validated**.		• The **Spectrum validation filter** selects from your data only the results that match your setting. • In this case, it selects only the spectra that do not yet have a search result that you have designated as valid.

Steps	Detailed Instructions	Comments
4 Set the **Validation preset** to **good-spectrum.**		When you process your own data, you have three choices: • If in the next step you set the **Filter by** parameter such that poor spectra are likely to be excluded, select **good-spectrum**. • Otherwise, select **bad-spectrum.** • Select **reset** if you want to undo a previous designation.
5 Set **Filter by**.	• For this exercise, keep the **Maximum tag length** option and set a filter of **>7**.	• With the **Maximum tag length** filter, higher numbers correlate with spectra of better quality. The setting of **7** filters and keeps only the best spectra. • The software has a number of additional filters. These include intensity and number of b/y pairs.
6 Select **Review Fields**.	**a** Keep the default settings. **b** Check to see that your settings look like those in Figure 24 on page 74.	
7 Click the **Summarize** button.	**a** Check that your results are similar to those in Figure 25 on page 75. Check that your results show: ◦ A table at the top of the page ◦ The Spectrum Viewer with the first spectrum at the bottom of the page	• Your results may vary slightly. • Note that text near the top of the summary table (just under the blue buttons) indicates that you have displayed about 80 potentially good spectra.

Steps	Detailed Instructions	Comments
8 Use the Spectrum Viewer to determine how much potential fragmentation information is contained in the spectrum.		• The Spectrum Viewer shows your extracted MS/MS spectrum annotated with an amino acid sequence that could explain the mass gaps between the ions. • In blue and red, you see potential b and y ions. The other ions are in black. • This is not intended to be a *de novo* interpretation, but merely shows how much plausible sequence information may be contained in the spectrum. • To learn more about the Spectrum Viewer, see "Exercise 10. Examine results with the Spectrum Viewer" on page 44.
9 Categorize the spectrum.	• For this exercise, keep the **good-spectrum** designation.	• If you wanted to change the designation to bad-spectrum, you would use one of two methods: ◦ Select **bad-spectrum** from the drop-down list under the **Validation category** heading. ◦ Click the **bad-spectrum** button near the top of the page.

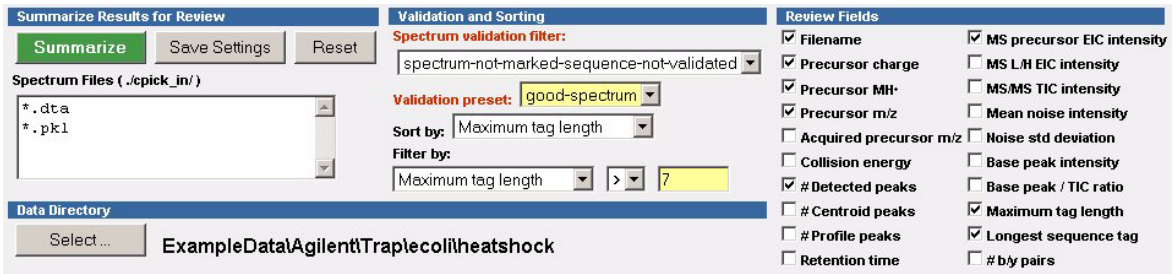


Figure 24 Spectrum Summary settings

Agilent Spectrum Mill - Spectrum Summary
Spectrum Mill
Summary Settings
Protein/Peptide Summary
Build TIC
MS/MS Search
Tool Belt
Help
Down
Up
HEC_0020.1339.1349.2.pkl
1
of 83
Update
reset
good-spectrum
bad-spectrum
Click to set category of current spectrum
Data Directory: msdataSM\ExampleData\\Agilent\\Trap\\ecoli\\heatshock
Validation Preset: spectrum-not-marked-sequence-not-validated Input files: 83 - 83 files (after filtering by maxSequenceTagLength > 7)
Validation category
#
Filename
z
Precursor MH+
Precursor m/z
Detected Peaks
MS Precursor EIC Intensity
Max Tag Length
Longest Sequence Tag
good-spectrum 1 HEC_0020.1339.1349.2 2 2164.0400 1082.5200 25 9.11e+007 11 [493.17]MNVANWVVNLH[391.23]
good-spectrum 2 HEC_0020.1708.1711.2 2 2163.3700 1082.1900 25 2.60e+007 11 [391.32]LFSVVSVSVVN[640.67]
good-spectrum 3 HEC_0020.2691.2694.2 2 2162.9400 1081.9700 25 7.09e+006 11 [391.26]LFSVANWVVNL[527.13]
good-spectrum 4 HEC_0040.1232.1241.2 2 2164.1700 1082.5900 25 6.13e+007 11 [510.09]NVNANWVVNLH[391.46]
good-spectrum 5 HEC_0020.1294.1298.2 2 2163.2800 1082.1400 25 1.23e+008 10 [391.49]LFSVANWVVN[640.48]
good-spectrum 6 HEC_0020.1318.1321.2 2 2164.2400 1082.6200 25 9.01e+007 10 [391.34]LFSVANWVVN[641.59]
good-spectrum 7 HEC_0020.1387.1391.2 2 2163.1500 1082.0800 25 4.53e+007 10 [391.38]LFSVANWVVN[640.44]
good-spectrum 8 HEC_0020.1451.1453.2 2 2163.8800 1082.4400 25 3.13e+007 10 [391.24]LFSVANWVVN[641.20]
good-spectrum 9 HEC_0020.1496.1496.0 2 2028.9400 1014.9700 25 1.62e+007 10 [407.24]LNGARDGTAL[651.23]
good-spectrum 10 HEC_0020.1537.1549.2 2 2163.0400 1082.0200 25 4.77e+007 10 [391.33]LFSVQGWVVN[640.39]
good-spectrum 11 HEC_0020.1556.1564.2 2 2163.1100 1082.0600 25 3.11e+007 10 [391.34]LFSVANWVVN[640.41]
good-spectrum 12 HEC_0020.1595.1609.2 2 2162.8700 1081.9400 25 3.98e+007 10 [391.33]LFSVANWVVN[640.21]
good-spectrum 13 HEC_0020.1667.1681.2 2 2164.0800 1082.5400 25 5.26e+007 10 [391.29]LFSVANWVVN[641.44]
MSTag 1
[493.17]
M N V A N W V V N L H
[391.23]
Reset
Labels
Grow^^
Grow
Shrink
Full
All/b-y
All/c-z
8.27e+5
10.0%
P(m/z)
P(m/z)-NH3
HEC_0020.1339.1349.2.pkl
Mass (m/z)
MH+: 2164.0327 m/z: 1082.52 z: 2
y b b-H2O y-H2O a b-NH3 y-NH3 y++ y(++)-H2O b++ y+3 b+3 b+H2O y-H3PO4 y-2H3PO4 b-H3PO4
[256.93]LVGEDVHNImT[362.24]
Go
Bars None
I-b I-b-H2O I-b-NH3 I-a c c' z' z'' c++ c'++ z'++ z''++
Rank
<-
->

Figure 25 Spectrum Summary results

Steps	Detailed Instructions	Comments
10 Display the next spectrum.	• Do one of the following: ◦ Click the **Down** button at the top of the page to display the next spectrum at the bottom of the page. ◦ In the summary report, click the number **2** under the **#** header. Again, this displays the spectrum at the bottom of the page. ◦ In the summary report, click the second entry under the **Longest Sequence Tag** header. This displays the Spectrum Viewer in a new window.	
11 Examine and categorize the spectrum.	• For this exercise, keep the **good-spectrum** designation.	
12 Repeat steps 11 and 12 for as many spectra as you care to review.		
13 Click the green **Update** button.		• This makes your designations a permanent part of the data record. • You have now designated approximately 80 spectra as good. • Your results may vary slightly.
14 (Optional) Vary input parameters and investigate the results.	**a** To change parameters and create a new summary table, click the **Summary Settings** button to start over. **b** For this exercise, do not click the **Update** button again. If you do, your results will not match in the next exercise.	

Exercise 5. Perform Sherenga *de novo* sequencing

Now that you have found some additional good spectra, you decide to try to identify them. If these were your samples, you might search another database in identity mode, conduct additional variable modifications or homology searches, or try *de novo* sequencing. In this exercise, you use Sherenga *de novo* sequencing to gain insight into possible amino acid sequences for these peptides. For more information about *de novo* sequencing, see the Sherenga *de novo* Sequencing chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Steps	Detailed Instructions	Comments
1 Navigate to the Sherenga *de novo* Sequencing page.	• From the Spectrum Mill home page, click the link to the *de novo* Sequencing page.	
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just been working with this data, the data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Set the **Validation filter** to **good-spectrum-sequence-not-validated.**		• The **Validation filter** selects from your data only the results that match your setting. • In this case, it selects only the spectra that you just designated as "good" with the Spectrum Summary page.

Steps	Detailed Instructions	Comments
4 Set the other parameters as shown in Figure 26 on page 79.	• Keep the default settings.	• The default settings are shown in the online help. To access the online help, click the **Help** button for the form. • When you process your own data, be sure to properly set the **Scoring** parameter. For example, if you generated your data with a QTOF instrument or an ion trap in ETD or mixed ETD/CID mode, you would change the **Scoring** setting. The Sherenga algorithm adapts to fragmentation patterns from different instruments.
5 Start the sequencing.	**a** Click the **Sequence** button. **b** View sequencing progress in the **Results** area to the right of the Data Extractor form. See Figure 27 on page 79. **c** Scroll to the *top* of the **Results** area to see the message that indicates that sequencing is finished.	

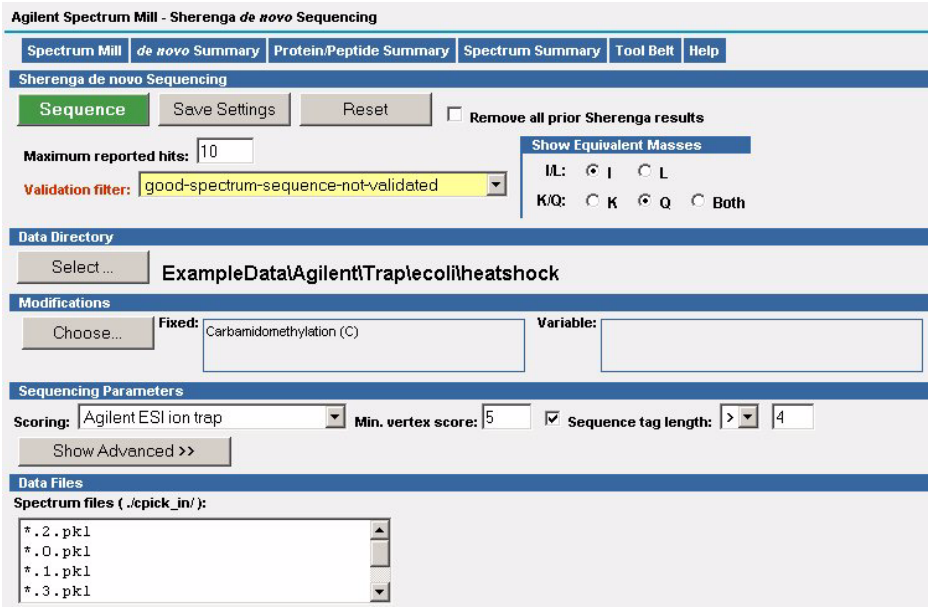


Figure 26 Sherenga *de novo* Sequencing settings

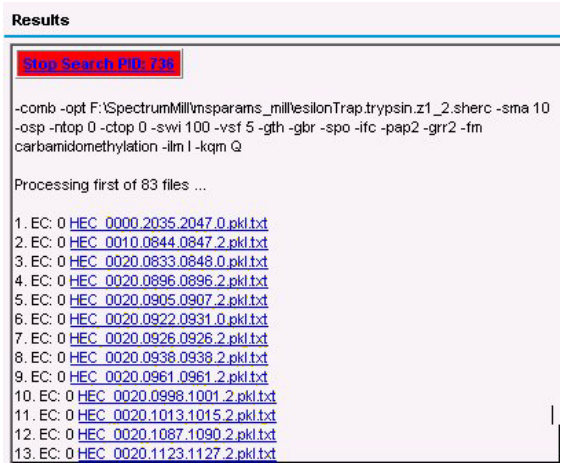

Results
Stop Search PID: 736
-comb -opt F:\SpectrumMill\nsparams_mill\esilonTrap.trypsin.z1_2.sherc -sma 10 -osp -ntop 0 -ctop 0 -swi 100 -vsf 5 -gth -gbr -spo -ifc -pap2 -grr2 -fm carbamidomethylation -ilm I -kqm Q
Processing first of 83 files ...
1. EC: 0 HEC_0000.2035.2047.0.pkl.txt
2. EC: 0 HEC_0010.0844.0847.2.pkl.txt
3. EC: 0 HEC_0020.0833.0848.0.pkl.txt
4. EC: 0 HEC_0020.0896.0896.2.pkl.txt
5. EC: 0 HEC_0020.0905.0907.2.pkl.txt
6. EC: 0 HEC_0020.0922.0931.0.pkl.txt
7. EC: 0 HEC_0020.0926.0926.2.pkl.txt
8. EC: 0 HEC_0020.0938.0938.2.pkl.txt
9. EC: 0 HEC_0020.0961.0961.2.pkl.txt
10. EC: 0 HEC_0020.0998.1001.2.pkl.txt
11. EC: 0 HEC_0020.1013.1015.2.pkl.txt
12. EC: 0 HEC_0020.1087.1090.2.pkl.txt
13. EC: 0 HEC_0020.1123.1127.2.pkl.txt

Figure 27 Display while Sherenga *de novo* Sequencing is in process

Exercise 6. Review results from Sherenga *de novo* sequencing

In this exercise, you review the *de novo* interpretations that you generated in the previous exercise.

Steps	Detailed Instructions	Comments
1 Navigate to the Sherenga *de novo* Summary page.	• From the Sherenga *de novo* Sequencing page or the Spectrum Mill home page, click the link to the *de novo* Summary page.	
2 Check that your **Data Directory** is set to **heatshock**.	• If you have just been working with this data, the data directory should already be set correctly. If not, click the **Select...** button to select the **heatshock** folder.	
3 Set the other parameters as shown in Figure 28 on page 81.	• Keep the default settings.	The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
4 Click the **Summarize** button.		
5 View the overall results table.	**a** Check to see that your results look like the Sherenga Summary results shown in Figure 29 on page 81. **b** Adjust the size of this section to see the results better. ◦ Rest your mouse over the page section split line. ◦ When the pointer becomes a double-sided arrow, drag the arrow to move the split line.	File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl.txt, where: • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)

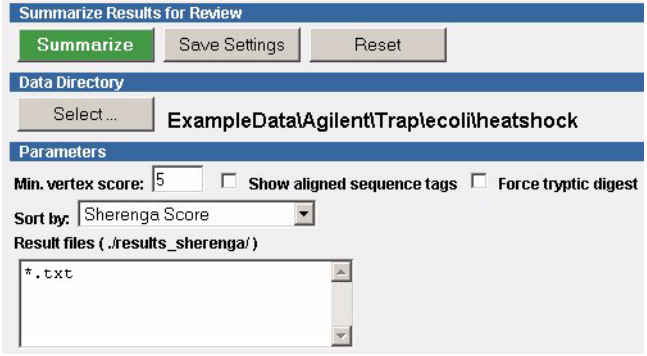


Figure 28 Settings for Sherenga Summary form

Results

SS	TL	TS	MS/MS Search Delta MH+	Filename	Sherenga Sequence Tag	MS/MS Search Result
307	10	5.4	0.7	HEC_0020.1556.1564.2.pkl.txt	[414.52]NINVVDAGGAGQFISA[215.27]	MGLVIKAALGALVVLLIGVLAK
292	10	4.2	0.7	HEC_0020.1766.1780.2.pkl.txt	[251.35]YNINVVDAGGVASFGG[372.49]	MGLVIKAALGALVVLLIGVLAK
285	10	4.1	0.7	HEC_0020.2376.2385.2.pkl.txt	[414.41]NINVVSGGVAGQFI[373.45]	MGLVIKAALGALVVLLIGVLAK
283	9			HEC_0020.2242.2256.2.pkl.txt	V[292.61]IFSVAAGTAVGVGIN[396.32]	
280	10	5.5	0.6	HEC_0020.1537.1549.2.pkl.txt	[414.38]NGGNVVDAGGEQFI[157.54][215.86]	MGLVIKAALGALVVLLIGVLAK
278	10	6.4	0.7	HEC_0020.1387.1391.2.pkl.txt	[327.27]SGGIGGVVPGSMGQAATS[303.99]	MGLVIKAALGALVVLLIGVLAK
276	10	4.7	0.9	HEC_0020.1294.1298.2.pkl.txt	[414.45]NGGNVVEGGGAGQFI[373.48]	MGLVIKAALGALVVLLIGVLAK
274	9			HEC_0020.1510.1510.0.pkl.txt	[324.48]ITIIATGGTIAGGGD[158.35][229.18]	
269	10			HEC_0020.1595.1609.2.pkl.txt	[414.34]NIPASGVSGGAGEFI[157.93][215.57]	
267	10	6.7	0.8	HEC_0020.1690.1702.2.pkl.txt	[414.38]NIGGVVGGAGGEGAFI[373.39]	MGLVIKAALGALVVLLIGVLAK
264	9	5.0	1.3	HEC_0020.1429.1444.2.pkl.txt	[414.49]NINVVADGGEGAFI[373.31]	MGLVIKAALGALVVLLIGVLAK
263	9	4.5	0.4	HEC_0040.1298.1311.2.pkl.txt	[212.87]IA[142.23]IQGDVAGEVGGHN[371.94]	MGLVIKAALGALVVLLIGVLAK
262	11	5.2	1.6	HEC_0020.1339.1349.2.pkl.txt	[396.35]GGNAGGVGGADVVNHI[157.85][215.56]	MGLVIKAALGALVVLLIGVLAK
262	9	6.6	1.1	HEC_0020.1583.1588.2.pkl.txt	[414.39]GGINVVADNEQFI[373.44]	MGLVIKAALGALVVLLIGVLAK
257	11	4.8	1.0	HEC_0020.1708.1711.2.pkl.txt	[414.50]NIGGVVSGGVEQFI[373.44]	MGLVIKAALGALVVLLIGVLAK
247	11			HEC_0020.2691.2694.2.pkl.txt	[215.06][199.28]NIGGVTGAGGGSVAFI[373.26]	
246	10	3.5	1.0	HEC_0020.2223.2232.0.pkl.txt	[235.61][172.08]DSSSAVDAGGVAPNH[373.37]	MGLVIKAALGALVVLLIGVLAK
245	9			HEC_0020.2448.2460.2.pkl.txt	[414.46]PTGVSVDANAVPHI[373.50]	
245	10	5.0	0.5	HEC_0040.1280.1283.2.pkl.txt	[414.45]NIGGVVMGGGSGG[184.87]M[373.28]	MGLVIKAALGALVVLLIGVLAK

Figure 29 Sherenga Summary results

Steps	Detailed Instructions	Comments
6 Review detailed results for a single spectrum.	**a** Click a link under the **Filename** header. Check that you see: • The Spectrum Viewer, as shown in Figure 30 on page 83. (You may need to move the page section split line again.) • Detailed Sherenga results, as shown in Figure 31 on page 83. **b** If you see an **MSTag 1** sequence at the top of the Spectrum Viewer, click the **Sherenga 1** sequence and make sure the sequence turns from gray to white background. If you do not see an **MSTag1** sequence, skip this step. **c** Click the **Rank** arrow buttons, shown in the lower right-hand corner of Figure 30 on page 83, to navigate through the various Sherenga interpretations. Note that many of the interpretations are similar. **d** To display your own sequence above the spectrum, type the sequence in the box and then click the **Go** button. 	• Note that the detailed Sherenga results list multiple Sherenga interpretations for the same spectrum. The most probable interpretation (Sherenga 1) is initially displayed at the top of the Spectrum Viewer. • If there are MS/MS Search results for these spectra, these are displayed along with the Sherenga interpretation. This makes it easy to compare the two. • To learn more about the Spectrum Viewer, see "Exercise 10. Examine results with the Spectrum Viewer" on page 44.
7 Repeat step 6 for as many spectra as you care to view.		• While *de novo* sequencing gives you valuable insight into possible spectral interpretations, you do not validate these results, and they are not included in final results summaries. • However, the Sherenga results are saved and can be reviewed at any time from the Sherenga *de novo* Summary page.

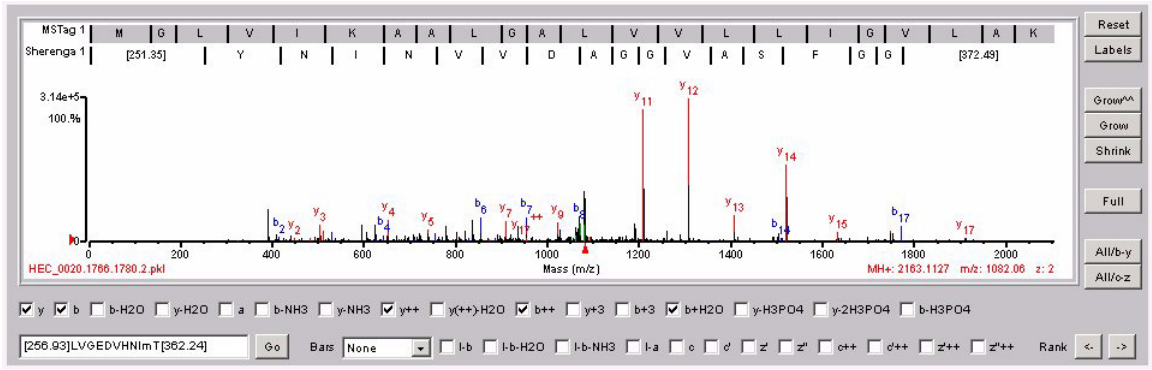


Figure 30 Spectrum Viewer with Sherenga results

ExampleData\Agilent\Trap\ecoli\heatshock HEC_0020.1766.1780.2.pkl.txt HEC_0020.1766.1780.2.pkl.spo

Rank	Orig. Score	Score	MS-Tag Sequence	Sherenga Sequence Tag	Matched Precursor MH+
		4.23	MGLVIKAALGALVVLLIGVLAK	[251.35]YNINVVDAGGVASFGG[372.49]	2162.4073
1.	292.0	292.0	[251.35] Y N I N V V D A G G V A S F G G [372.49]		2163.11
2.	291.1	291.1	[251.35] Y N I N V V D A G G V A S F I [158.15] [215.27]		
3.	290.5	290.5	[251.35] Y N I N V V D A G G E Q F G G [372.49]		
4.	289.9	289.9	[414.35] N I N V V D A G G V A S F I [158.15] [215.27]		
5.	289.7	289.7	[251.35] Y N I N V V D A G G V A S F I [373.42]		
6.	289.6	289.6	[251.35] Y N I N V V D A G G E Q F I [158.15] [215.27]		
7.	288.5	288.5	[414.35] N I N V V D A G G E Q F I [158.15] [215.27]		
8.	288.3	288.3	[251.35] Y N I N V V D A G G E Q F I [373.42]		
9.	288.1	288.1	[251.35] Y N I R G V D A G G V A S F G G [372.49]		
10.	287.7	287.7	[251.35] Y N I N V V D A G G A V S F G G [372.49]		

Figure 31 Detailed results from Sherenga *de novo* Sequencing

3
Processing Ion Trap - ETD Data

Exercise 1. (Optional) Transfer ETD data file to the Spectrum Mill server 86
Exercise 2. Extract the spectra 88
Exercise 3. Create indices for a user database 92
Exercise 4. Run database search 94
Exercise 5. Run Autovalidation 100
Exercise 6. Display valid database search results in peptide mode 103
Exercise 7. Examine results with the Spectrum Viewer 106
Exercise 8. Display valid database search results in protein mode 109

This chapter shows how to use the Spectrum Mill workbench to process electron transfer dissociation (ETD) data from an Agilent ion trap. The data file used in these exercises includes both collision-induced dissociation (CID) and ETD spectra.

Exercise 1. (Optional) Transfer ETD data file to the Spectrum Mill server

To process files that contain ETD spectra with the Spectrum Mill workbench, you first move or copy the files into the appropriate directory on the Spectrum Mill server.

To make it easy to compare data sets, it is important that you set up the appropriate directory structure for your files on the Spectrum Mill server. Whenever you want to compare samples in a set, you need to set up a subdirectory for each sample. This directory may contain data files from multiple sample fractions or gel slices.

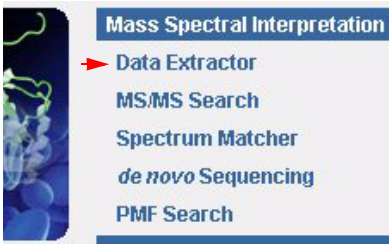
NOTE The ETD example file may have been installed on the server at the time the Spectrum Mill workbench was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

Steps	Detailed Instructions	Comments
1 Copy the example data file from the **X:\ExampleData\msdataSM\ ExampleData\Agilent\Trap\ETD** folder on the *Spectrum Mill Example Data CD* onto your client PC.	• Copy the **ETD** folder from the *Spectrum Mill Example Data CD* to any folder on your client PC.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • This exercise simulates the likely laboratory scenario of file transfer from a client or instrument PC to the server; you may transfer files directly from the CD to the server if you prefer.

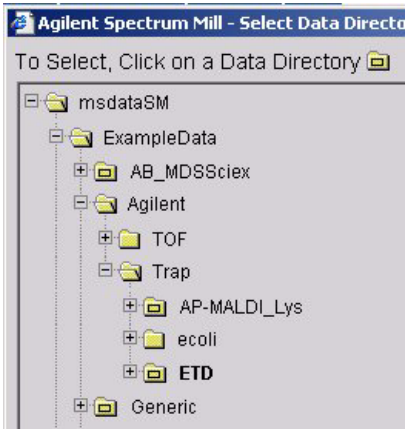
Steps	Detailed Instructions	Comments
2 Create the following folder on the Spectrum Mill server: **msdataSM\ExampleData\Agilent\Trap\ETD.**	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In **msdataSM**, create a new folder for the example file. Create **msdataSM\ExampleData\Agilent\Trap\ETD.** **d** You need only one folder because this exercise uses a single sample.	• If you don't know how to find your Spectrum Mill file system, ask the person who installed your software. • Do not include spaces or parentheses in your folder name. • When you process your own samples, remember to set up a separate folder for each sample. Each folder should contain all sample fractions. • You may create up to ten folders between **msdataSM** and your data files, but shorter path lengths reduce memory usage, especially for large data sets.
3 Copy the example file to the new folder.	• Copy or move **CID_ETD_Mix.D** from your client PC to the new folder on the server PC.	When you process your own data, remove any spaces or parentheses in the data file names. For example, change the file **my sample.d** to **my_sample.d.**
4 Make sure you have both read and write permissions for the data folder you just created on the server.	**a** Right-click the **ETD** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked. **c** In the **Confirm Attribute Changes** dialog, click **Apply changes to this folder, subfolders, and files**. **d** Click **OK**. **e** If necessary, repeat step **a**. **f** Click the **Security** tab. **g** Make sure all user groups have full permissions. **h** Click **OK**.	If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors.

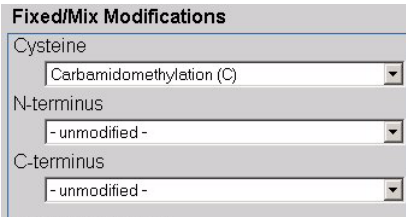
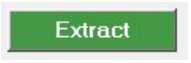
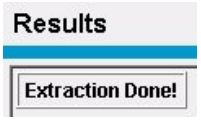
Exercise 2. Extract the spectra

The Spectrum Mill Data Extractor preprocesses raw data files to extract high-quality spectra for database searches. When data files contain spectra from both ETD and CID, the Data Extractor extracts and merges nearby MS/MS spectra of the same dissociation type. It does *not* merge ETD with CID spectra. For each extracted spectrum, the Data Extractor specifies the dissociation method in a file that is subsequently accessed by the MS/MS Search program. This is important since ETD and CID spectra show different fragmentation patterns and the scoring rules for MS/MS Search are different.

Steps	Detailed Instructions	Comments
1 Start the Spectrum Mill workbench on your PC.	• Double-click the desktop icon to launch the Spectrum Mill workbench.	You see the Spectrum Mill home page.
2 If you don't see the Spectrum Mill icon, launch your web browser and type the URL for the Spectrum Mill home page.	**a** Click your Internet Explorer icon. **b** In the web browser window, type the URL for the Spectrum Mill server and press **Enter**.	If you don't know the URL, ask the person who installed the software.
3 Check that you see the Spectrum Mill home page.		
4 If you don't, check that the server is booted up and that your URL is correct.		
5 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link. 	

Steps	Detailed Instructions	Comments

6 Select the **msdataSM\ ExampleData\Agilent\Trap\ETD** data directory.	**a** Click the **Select...** button near the middle of the form.
	
	b Expand the directory tree and click the word **ETD** to select that data file directory.
	 
	c Make sure that the name of the **ETD** directory changes to a **bold** font. This indicates that it has been selected.
	d Click **OK.** The data directory appears as in Figure 32 on page 91.
		• Directories are identified by different types of icons:
		  ◦ Plain folders indicate directories that do not have data files directly beneath them.
		  ◦ Folders with rectangles indicate data directories.
		  ◦ Folders with line spectra (bar graphs) indicate data files.
		• Try clicking the names of each type to see which turn **bold**, indicating that they are selectable.
		• The software remembers your data file selection when you go to other Spectrum Mill forms.
		• If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
7 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **Carbamidomethylation**.  **c** Click **OK.** The name of the modification appears as in Figure 32 on page 91.	• Carbamidomethylation is the default when you first install the Spectrum Mill workbench, so you may be able to skip this step. • To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
8 Set other parameters as shown in Figure 32 on page 91.	**a** Keep the defaults whenever appropriate. This example uses all default settings, except for **Data Directory**. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	
9 Start the extraction.	**a** Click the **Extract** button.  **b** View extraction progress in the **Results** area to the right of the Data Extractor form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that extraction is finished. 	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction is taking place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

Steps	Detailed Instructions	Comments
10 (Optional) View the **cpick_in** subdirectory to see the files Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\Agilent\Trap\ETD\cpick_in**. **b** Notice the new files created there. Each one represents an extracted spectrum from your raw data file.	File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)

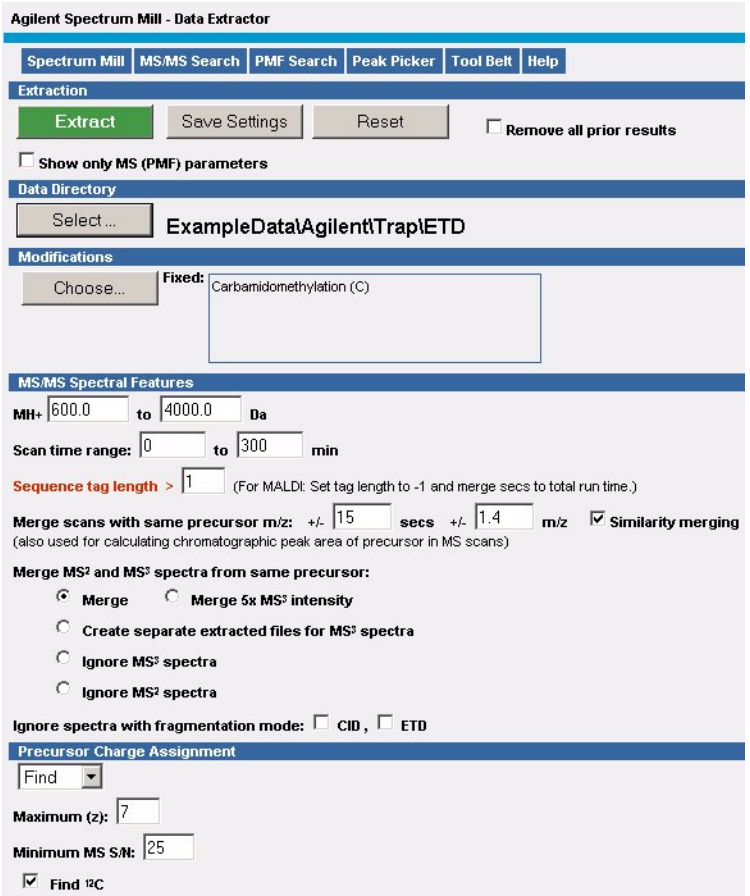


Figure 32 Settings for Data Extractor

Exercise 3. Create indices for a user database

The Spectrum Mill workbench allows you to create and search user databases. In this exercise, you create indices for a user database so you can later search it.

Steps	Detailed Instructions	Comments
1 If necessary, copy the user database **SwissProt.CID_ETD_Mix** to your Spectrum Mill server.	**a** If the **SwissProt.CID_ETD_Mix** database is already installed in your Spectrum Mill database directory, skip to step 2. **b** If not, find the file **X:\Example Databases\SwissProt.CID_ETD_Mix** on your *Spectrum Mill Installation CD*. **c** Copy it to your database directory on your Spectrum Mill server.	• The database directory on the server likely has a name like **X:\seqDB**. • If you are not sure, ask your system administrator.
2 Make sure security permissions are set properly for the folder that contains the file **dbname.js**.	**a** Right-click the **SpectrumMill\millhtml\SM_js** folder and select **Properties**. **b** Click the **Security** tab. **c** Make sure all user groups have full permissions. **d** Click **OK**.	If the Spectrum Mill workbench cannot write to this folder, you may encounter errors.
3 Navigate to the Protein Databases page.	• From the Spectrum Mill home page, click the Protein Databases link. 	
4 Fill in the form as shown in Figure 33 on page 93.	**a** On the Protein Databases page, click the **Create indices for new database** option. (This is the default.) **b** Set **Newly downloaded database** to **SwissProt.CID_ETD_Mix.**	For **Newly downloaded database**, you type the exact database file name.

Steps	Detailed Instructions	Comments
5 Click the **Create Indices** button.	• After a short wait, check the bottom of the page for text that indicates that the software is creating the database indices.	These indices bear no resemblance to those used by another popular database search program.

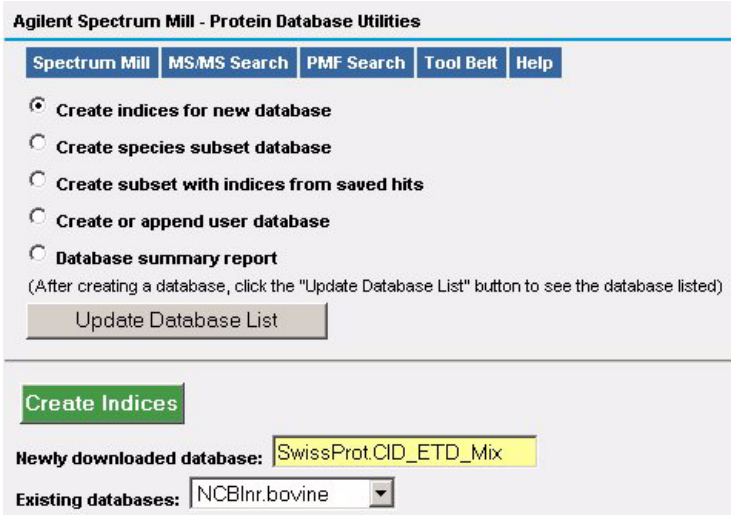


Figure 33 Create indices for user database

Exercise 4. Run database search

After you have extracted your spectra, you are ready to search each spectrum against a protein or DNA database. As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal of identifying as many spectra as possible. These exercises illustrate only a single identity mode search. In identity mode, the spectra must be consistent with the database sequence. When you process your own data, you might also search in variable modifications or homology mode, where spectra may show modifications relative to the database sequence.

The MS/MS Search module applies the appropriate rules and scoring for ETD and CID spectra. It searches ETD and CID spectra in separate batches.

Steps	Detailed Instructions	Comments
1 Navigate to the MS/MS Search page.	• Do one of the following: ◦ From the Protein Databases page, click the **MS/MS Search** button. ◦ From the Spectrum Mill home page, click the link to MS/MS Search. 	Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Data Extractor page to the MS/MS Search page.
2 Check that your **Data Directory** is set to **ETD**.	• If you have just performed the data extraction, your **ETD** data directory should already be set correctly. If not, click the **Select...** button to select the **ETD** folder.	

Steps	Detailed Instructions	Comments
3 Select the **SwissProt.CID_ETD_Mix** database.	• For **Database** (under **Search Parameters**) select the **SwissProt.CID_ETD_Mix** database.	If the database name does not appear, reload or refresh the MS/MS search form.
4 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **Carbamidomethylation** **c** Click **OK.** The name of the modification appears as in Figure 34 on page 99.	• The modification is likely already set. • Three types of modifications are available for MS/MS Search: ◦ The **fixed** modifications are assumed to apply universally and are searched in a single search cycle. ◦ The **mix** modifications trigger cyclic MS/MS searches, where a different form of the modification is searched in each cycle. ◦ The **variable** modifications allow for both modified and unmodified forms within a peptide. All variable modifications are searched in each cycle. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.

Steps	Detailed Instructions	Comments

5 Set up the **Matching Tolerances**.	**a** Change the **Instrument** setting to **Agilent ESI ion trap MIX-CID-ETD**.
	b Leave the other settings at default values.
		• When you set **Instrument** to **Agilent ESI ion trap MIX-CID-ETD**, the software applies the appropriate scoring for CID and ETD spectra.
		• If your data file contained only ETD spectra, you would set **Instrument** to **Agilent ESI ion trap ETD**.
		• If you have spectra with higher charge states, change the setting for **Maximum ambiguous precursor charge**.
		  ◦ For enzymes such as trypsin that produce shorter peptides, it is best to try **3** first, then see if you get more valid hits by setting to **4** or **5**.
		  ◦ For enzymes such as LysC that produce longer peptides, it is best to try **4** first, then try **5** in a subsequent search.

Steps	Detailed Instructions	Comments
6 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Proton mobility scoring**. **c** Mark the check box for **Dynamic peak thresholding**. **d** Verify that the **Search mode** is set to **Identity**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Protein mobility scoring** is a scoring enhancement for ion trap data. It applies a bonus or penalty based on the expected relative intensities for MS/MS fragments from a given peptide sequence. When you process your own ion trap data, it is best to use this scoring, except with modifications that significantly alter the fragmentation patterns of peptides. • The software selectively applies proton mobility scoring to CID spectra and does not apply it to ETD spectra. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.

Steps	Detailed Instructions	Comments
7 Make sure other parameters are set as shown in Figure 34 on page 99.	**a** Keep the defaults whenever appropriate. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	
8 Start the search.	**a** Click the **Start Search** button **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • CID and ETD data are searched in separate sets of batches, as you can see (in a few minutes) in the **Results** area of the page. • Search time varies depending on the size of the database searched. This search goes fairly quickly because you search a small user database. • You can use your client PC for other tasks while the search is taking place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

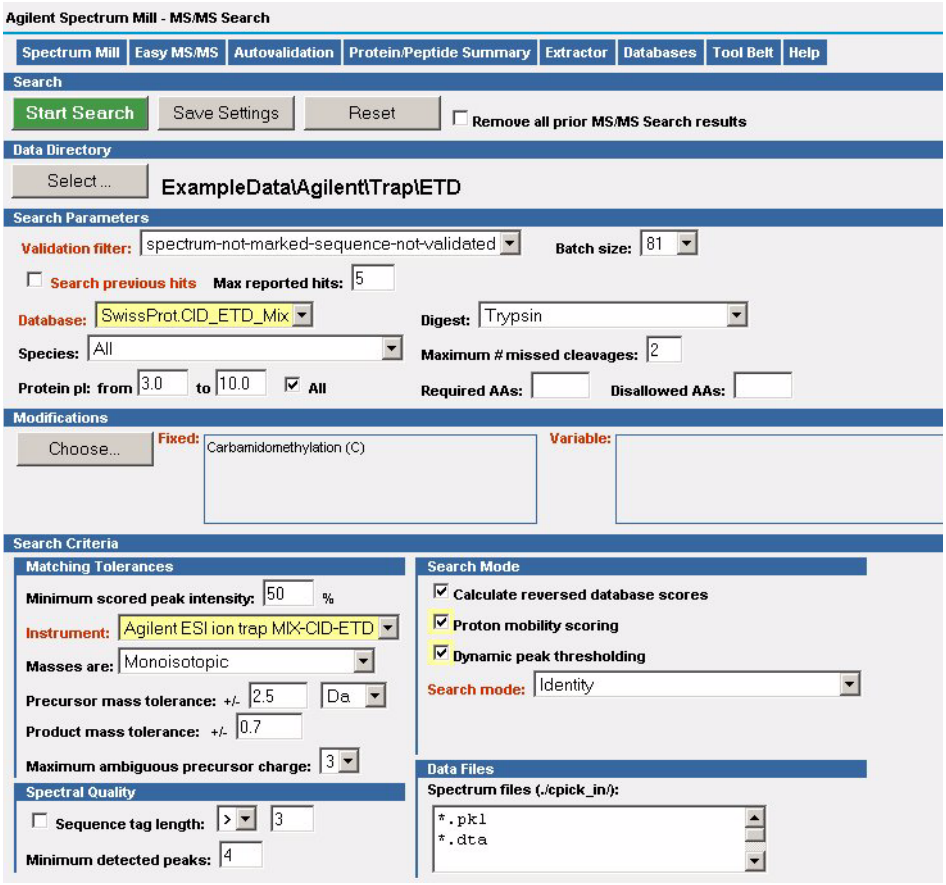


Figure 34 MS/MS Search settings

Exercise 5. Run Autovalidation

After you have completed a database search, you validate the good results. Validation means that you accept that the matches are correct.

The Spectrum Mill workbench provides a means for segregating search results that contain a valid interpretation of an MS/MS spectrum from those that do not. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

The Spectrum Mill workbench provides two ways of validating results. One way uses the Autovalidation page, and is totally automated. You use this method only to validate the highest-scoring results — those that do not require manual review. The other method of validating uses the Protein/Peptide Summary page for manual review and validation. This exercise describes autovalidation. To learn more about manual validation, see the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 35 on page 102.

2 Check that your **Data Directory** is set to **ETD**.	• If you have just performed data extraction and MS/MS search for this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **ETD** folder.	
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Keep the default scoring presets. **c** Click the **Validate Files** button. **d** Watch for a Spectrum Mill Validation Summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein. • Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20.

Steps	Detailed Instructions	Comments
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Keep the default scoring presets. **c** Click the **Validate Files** button. **d** Watch for a Spectrum Mill Validation Summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by peptide. Even if it finds only a single peptide for a protein, it validates the corresponding search results provided that the peptide score is high enough. • Using the default scoring, individual peptides must have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein.
5 Close the Autovalidation form.		

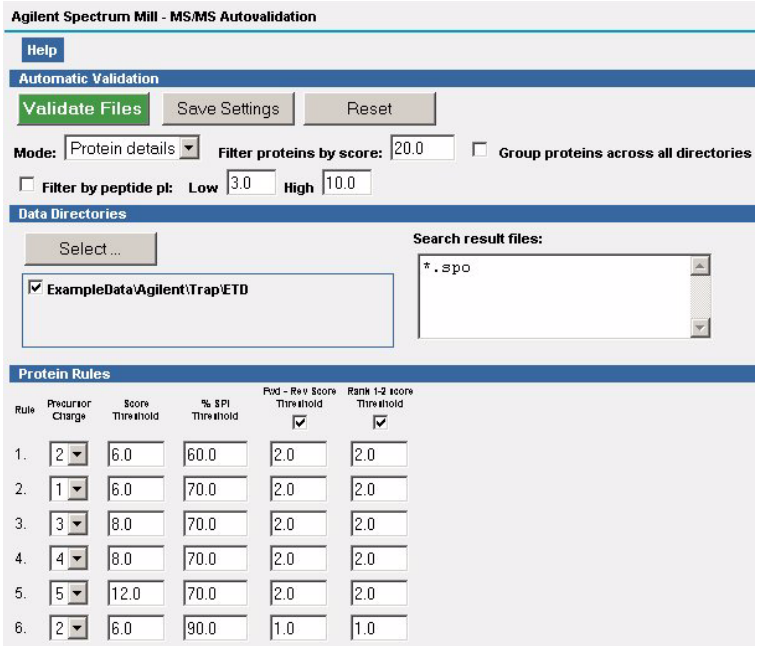


Figure 35 Autovalidation form

Exercise 6. Display valid database search results in peptide mode

In this exercise, you summarize the results that you just automatically validated. This summary uses peptide mode so that you can distinguish the ETD and CID results.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• Do one of the following: ◦ From the MS/MS Search page, click the **Protein/Peptide Summary** button. ◦ From the Spectrum Mill home page, click the link to Protein/Peptide Summary.	There are links to the Protein/Peptide Summary page from many other pages.
2 Check that your **Data Directory** is set to **ETD**.	• If you have just performed autovalidation for this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **ETD** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set the **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Set the **Validation preset** to **none**.		

Steps	Detailed Instructions	Comments
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	By setting to zero, you ensure that all valid data are displayed, regardless of score.
8 Set the **Review Fields**.	**a** Mark the check box for **Rank 1-2 score**. **b** Mark the check box for **Fragmentation mode**. **c** For others, keep the default settings. **d** Check to see that your form looks like that in Figure 36 on page 105. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		
10 Examine the summary report.	**a** Check that your results are similar to those in Figure 37 on page 105. **b** Note that each spectrum is marked with the fragmentation mode. Since ETD and CID spectra show different fragmentation patterns (see the *Application Guide*), consider the fragmentation mode as you review the data.	File names are in the format: Data_File_Name.aaaa.bbbb.c, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)

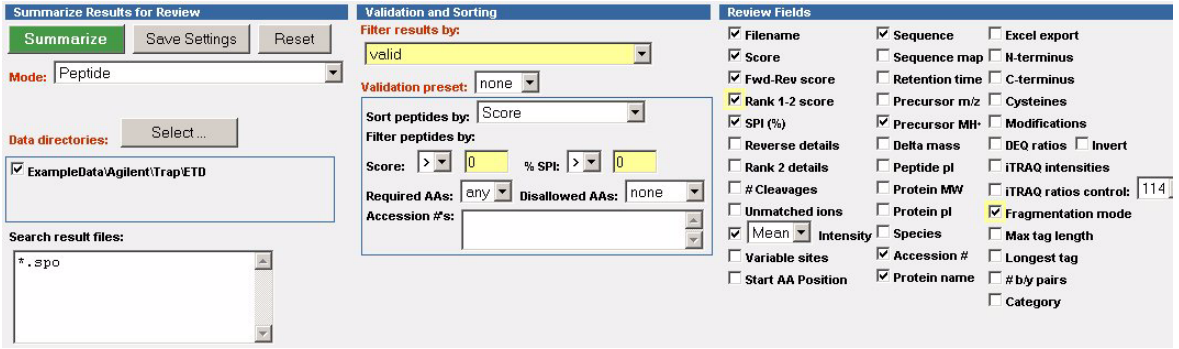


Figure 36 Settings to display valid results in peptide mode

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ETD
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 168 **Hits:** 168

#	Filename	z	Frag Mode	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	Spectrum Intensity	Sequence	MH+ Matched (Da)	Accession #	Protein Name
1	CID_ETD_Mix.1776.1776.3	3	CID	24.95	24.95	24.95	96.3	1.15e+008	(K)SYELPDGQVITIGNER(F)	1790.892	P02571	Actin, cytoplasmic 2 (Gamma-actin).
2	CID_ETD_Mix.1773.1773.0	2	CID	23.73	23.73	23.73	97.8	2.85e+008	(K)SYELPDGQVITIGNER(F)	1790.892	P02571	Actin, cytoplasmic 2 (Gamma-actin).
3	CID_ETD_Mix.1197.1197.2	2	CID	23.28	23.28	23.28	97.6	1.14e+008	(R)GLAGVENVTELK(K)	1229.674	P00489	Glycogen phosphorylase, muscle form (EC 2
4	CID_ETD_Mix.1295.1295.0	2	CID	22.32	22.32	22.32	97.6	1.10e+008	(R)IEELEEEIEAER(A)	1488.706	P02562	Myosin heavy chain, skeletal muscle (Fragme
5	CID_ETD_Mix.0707.0707.2	2	CID	21.90	21.90	21.90	97.4	1.85e+008	(K)TGQAPGFSYTDANK(N)	1456.670	P00006	Cytochrome c.
6	CID_ETD_Mix.1400.1400.2	2	CID	21.50	21.50	21.50	94.9	5.28e+007	(K)LEQQVDDLEGSLEQEK(K)	1859.887	P02562	Myosin heavy chain, skeletal muscle (Fragme
7	CID_ETD_Mix.1757.1757.2	2	CID	21.00	21.00	21.00	93.3	4.05e+008	(K)LLSYVDDEAFIR(D)	1440.737	P00489	Glycogen phosphorylase, muscle form (EC 2
8	CID_ETD_Mix.0182.0182.0	2	CID	19.84	19.84	19.84	91.7	2.24e+008	(K)VGDANPALQK(V)	1012.542	P00921	Carbonic anhydrase II (EC 4.2.1.1) (Carbonat
9	CID_ETD_Mix.1870.1870.0	2	CID	19.47	19.47	19.47	95.2	1.03e+008	(R)LSFNPTQLEEQCHI(-)	1715.806	P02754	Beta-lactoglobulin precursor (Beta-LG) (Aller
10	CID_ETD_Mix.0177.0177.0	2	CID	19.21	19.21	19.21	92.4	6.28e+007	(K)DSYVGDEAQSK(R)	1198.522	P02571	Actin, cytoplasmic 2 (Gamma-actin).
11	CID_ETD_Mix.1647.1647.0	3	CID	19.20	19.20	19.20	92.9	2.91e+007	(K)TKLEQQVDDLEGSLEQEK(K)	2089.030	P02562	Myosin heavy chain, skeletal muscle (Fragme
12	CID_ETD_Mix.1822.1863.2	2	CID	19.15	19.15	19.15	90.3	4.13e+008	(R)IGEEYISDLDQLR(K)	1550.770	P00489	Glycogen phosphorylase, muscle form (EC 2
13	CID_ETD_Mix.1352.1352.0	3	ETD	19.09	19.09	19.09	94.6	1.72e+008	(K)SLHTLFGDELCK(V)	1419.694	P02769	Serum albumin precursor (Allergen Bos d 6).
14	CID_ETD_Mix.1421.1421.0	2	CID	18.71	18.71	18.71	86.4	9.04e+007	(K)VPQVSTPTLVEVSR(S)	1511.843	P02769	Serum albumin precursor (Allergen Bos d 6).
15	CID_ETD_Mix.1734.1734.0	3	CID	18.51	18.51	18.51	84.0	1.74e+007	(K)ARPEFTLPVHFYGR(V)	1689.886	P00489	Glycogen phosphorylase, muscle form (EC 2
16	CID_ETD_Mix.1386.1386.2	2	CID	18.32	18.32	10.72	95.2	1.53e+008	(K)LVNELTEFAK(T)	1163.631	P02769	Serum albumin precursor (Allergen Bos d 6).
17	CID_ETD_Mix.1087.1087.2	2	CID	18.20	18.20	18.20	81.1	4.08e+008	(R)VDEDQPFPAVPK(W)	1341.669	P00722	Beta-galactosidase (EC 3.2.1.23) (Lactase).
18	CID_ETD_Mix.0617.0617.3	3	ETD	18.11	18.11	18.11	91.6	1.51e+008	(K)NFSDVHPEYGSR(I)	1407.629	P00432	Catalase (EC 1.11.1.6).
19	CID_ETD_Mix.0588.0588.2	2	CID	18.10	18.10	18.10	90.1	1.11e+008	(R)FNSANDDNVTQVR(T)	1479.682	P00432	Catalase (EC 1.11.1.6).
20	CID_ETD_Mix.1122.1122.0	3	CID	17.83	10.64	17.83	95.9	8.25e+007	(K)TGPNLHGLFGR(K)	1168.622	P00006	Cytochrome c.
21	CID_ETD_Mix.1669.1702.3	3	ETD	17.83	17.83	17.83	93.0	4.87e+008	(K)VWPHGDYPLIPVGK(L)	1577.848	P00432	Catalase (EC 1.11.1.6).
22	CID_ETD_Mix.1316.1316.2	2	CID	17.65	17.65	17.65	92.4	1.89e+008	(K)VFADYEEYVK(C)	1262.594	P00489	Glycogen phosphorylase, muscle form (EC 2
23	CID_ETD_Mix.1891.1891.3	3	CID	17.64	17.64	17.64	92.4	2.17e+007	(R)LSFNPTQLEEQCHI(-)	1715.806	P02754	Beta-lactoglobulin precursor (Beta-LG) (Aller
24	CID_ETD_Mix.1777.1777.3	3	ETD	17.47	17.47	17.47	89.4	1.15e+008	(K)SYELPDGQVITIGNER(F)	1790.892	P02571	Actin, cytoplasmic 2 (Gamma-actin).
25	CID_ETD_Mix.0652.0652.3	3	ETD	17.40	3.56	17.40	95.1	1.75e+008	(R)LFTGHPETLEK(F)	1271.663	P02188	Myoglobin.

Figure 37 Summary of valid peptides from autovalidation (partial listing)

Exercise 7. Examine results with the Spectrum Viewer

The Spectrum Viewer is a visual tool to evaluate MS/MS search results (Protein/Peptide Summary), to assess MS/MS spectral quality (Spectrum Summary), and to evaluate *de novo* sequencing results (Sherenga *de novo* Summary). This exercise focuses on use of the Spectrum Viewer for examination of MS/MS search results for ETD spectra. In this capacity, the Spectrum Viewer shows your extracted spectrum annotated with theoretical fragments from the top database search result. The viewer has flexible scaling and labeling options to allow you to better evaluate your results. For more comprehensive exercises on manual review of search results, see Chapter 1 of this guide.

Steps	Detailed Instructions	Comments
1 If you have not already done so, display a Protein/Peptide Summary report.	• Follow steps in "Exercise 6. Display valid database search results in peptide mode" on page 103.	
2 Display the first **ETD** spectrum.	**a** In the summary report, find the first spectrum where the **Frag Mode** is listed as **ETD**. **b** Click the number (under the **#** header) for that spectrum. **c** Check the bottom of the browser window, where you see the spectrum displayed in the Spectrum Viewer. **d** Check that your results are similar to those in Figure 38 on page 108.	

Steps	Detailed Instructions	Comments
3 Use the **All/c-z** button to change the spectrum annotations.	**a** Click the **All/c-z** button on the right side of the Spectrum Viewer. **b** Check that the peak labels change to show the c- and z-ions that are typical of ETD spectra. The labels also show precursor ions that have reduced charges. **c** Note that all check boxes for c- and z-ions are marked in the bottom row under the Spectrum Viewer. **d** Click the **All/c-z** button again. **e** Check that the peak labeling changes to show a subset of the c- and z-ions. **f** Note that a subset of check boxes for c- and z-ions are marked in the bottom row under the Spectrum Viewer.	• When you review a mixture of ETD and CID spectra, it is convenient to click the **All/c-z** button for ETD spectra and the **All/b-y** button for CID spectra. • The **All/b-y** button behaves similar to the **All/c-z** button, except it affects labelling of b- and y-ions. • In the spectrum, unmatched ions are shown in black, while matched ions are displayed in red and blue.
4 Select **Bars** options to locate all potential c- and z-ions.	**a** Find the **Bars** options at the bottom of the Spectrum Viewer. **b** Select **c** and observe the gray lines that indicate locations of all possible c-ions from the database match. **c** Check that your display looks similar to the one in Figure 39 on page 108. **d** Select **z'** and observe the gray lines that indicate locations of all possible z-ions from the database match. **e** Notice that the amino acid sequence reverses order when you switch from **c** to **z'**. **f** Select **None** to return to the default setting.	• By choosing **c** or **z'** for **Bars**, you quickly see how well the sequence from the summary table matches the spectrum. • Similarly, for CID spectra, choose **b** or **y** for **Bars**.

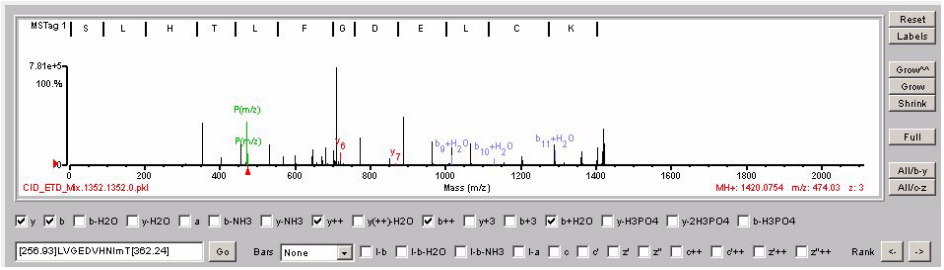


Figure 38 Spectrum Viewer - initial display

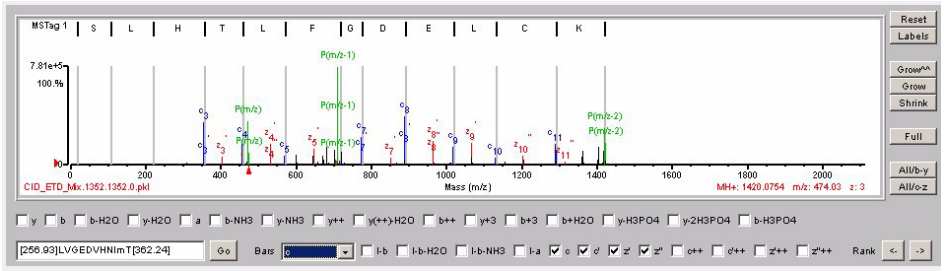


Figure 39 Spectrum Viewer, with annotations that help to evaluate ETD spectra

Exercise 8. Display valid database search results in protein mode

In this exercise, you summarize the validated results in protein mode.

Steps	Detailed Instructions	Comments

1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button at the top of the page.
	• If not, do one of the following:
	  ◦ From the MS/MS Search page, click the **Protein/Peptide Summary** button.
	  ◦ From the Spectrum Mill home page, click the link to Protein/Peptide Summary.
		• There are links to the Protein/Peptide Summary page from many other pages.
		• The **Summary Settings** button replaces the Spectrum Viewer with the form that shows the settings to summarize results.

2 Check that your **Data Directory** is set to **ETD**.	• If you have just examined ETD results in the peptide mode, your data directory should already be set correctly. If not, click the **Select...** button to select the **ETD** folder.	
3 Set the **Mode** to **Protein Summary**.		Notice that the summary input form changes to correspond with the new display mode.
4 Set the **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Leave **Sort proteins by** set to the default of **Score**.		• The **Sort** parameters determine how the data are sorted in the results display. • When you process your own data, select the setting that is most helpful to you.

Steps	Detailed Instructions	Comments
6 Set the **score** and **% SPI** filters to **> 0**.	**a** Set **Filter by protein score** to **> 0**. **b** For **Filter peptides by**, set both **Score** and **% SPI** to **> 0**.	When you set this to **> 0**, you are sure to display all the valid results.
7 Set the **Review Fields**.	**a** Mark the **Protein MW** check box. **b** Keep the other default settings. **c** Check to see that your form looks like that in Figure 40 on page 111. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The default settings are shown in the online help. Click the **Review Fields** blue dividing bar to access the online help.
8 Click the **Summarize** button.		
9 Examine the overall summary report.	• Check that your results are similar to those in Figure 41 on page 111.	Note the colored cells in the summary report. • The color-code indicates relative quantities. Dark red is highest, orange is intermediate, and yellow is lowest. • The **Distinct Peptides** column gives the number of peptides detected for each protein, • The **Mean Peptide Spectral Intensity** is an average of the intensities for all peptides detected for that protein. These intensities are calculated from extracted ion chromatograms from the precursor ions. • For more accurate quantitation, under **Review Fields**, for **Intensity**, select **Total** rather than **Mean**. • The intensity results are sufficient for studies where you are interested in differences of two-fold or more.

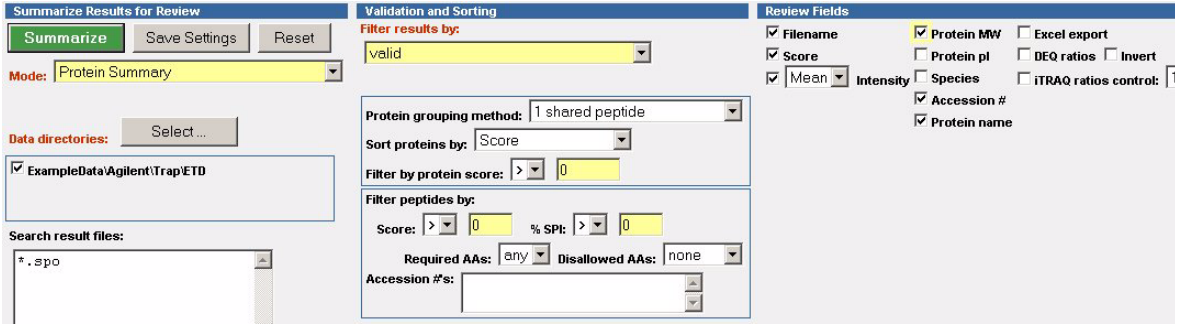


Figure 40 Settings to summarize valid results - Protein Summary mode

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Trap/ETD
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 168 **Hits:** 168
beginning to assemble proteins .... proteins assembled 0.034229 sec
proteins filtered by unique peptides 0.021053 sec
proteins filtered by score
calculated protein coverage maps 0.175316 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.020766 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 15 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Protein MW (Da)	Database Accession #	Protein Name
1	32	19	274.82	32	1.57e+008	69293.9	P02769	Serum albumin precursor (Allergen Bos d 6).
2	35	19	270.33	22	2.10e+008	97158.7	P00489	Glycogen phosphorylase, muscle form (EC 2.4.1.1) (
3	11	9	138.55	9	6.91e+007	125489.3	P02562	Myosin heavy chain, skeletal muscle (Fragments).
4	18	10	127.98	10	1.61e+008	116352.3	P00722	Beta-galactosidase (EC 3.2.1.23) (Lactase).
5	18	8	121.35	20	1.47e+008	57585.9	P00432	Catalase (EC 1.11.1.6).
6	15	9	111.35	13	1.72e+008	77753.7	Q29443	Serotransferrin precursor (Transferrin) (Siderophi
7	8	4	68.72	11	1.38e+008	41793.1	P02571	Actin, cytoplasmic 2 (Gamma-actin).
8	7	5	59.50	11	1.16e+008	49438.8	P00634	Alkaline phosphatase precursor (EC 3.1.3.1) (APASE
9	8	3	46.20	19	1.42e+008	19883.4	P02754	Beta-lactoglobulin precursor (Beta-LG) (Allergen B
10	4	2	39.73	24	1.35e+008	11572.4	P00006	Cytochrome c.
11	6	3	33.18	9	9.76e+007	35691.1	P46406	Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1
12	3	2	31.57	7	2.08e+008	28982.7	P00921	Carbonic anhydrase II (EC 4.2.1.1) (Carbonate dehy
13	1	1	17.40	7	1.75e+008	16951.6	P02188	Myoglobin.
14	1	1	15.26	3	2.81e+007	48057.8	P29952	Mannose-6-phosphate isomerase (EC 5.3.1.8) (Phosph

Figure 41 Protein summary of valid search results (partial listing)

4
Processing Q-TOF MS/MS Data

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 114
Exercise 2. Extract the spectra 116
Exercise 3. Search a database 121
Exercise 4. Autovalidate the hiqhest-quality results 125
Exercise 5. (Optional) Examine the validated results 129
Exercise 6. Display medium-quality results in preparation for manual validation 134
Exercise 7. Manually review and validate results 140
Exercise 8. Summarize valid results 142

This chapter provides familiarization exercises to process an example Agilent Q-TOF MS/MS data file with the Agilent Spectrum Mill MS Proteomics Workbench. In general, the steps to process Q-TOF MS/MS data are the same as those for ion trap MS/MS data. However, the MS/MS Search settings use a narrower mass tolerance, which often allows you to manually validate spectra with lower scores.

The peptide sample used in these familiarization exercises is a single fraction from the Agilent OFFGEL Fractionator, an electrophoretic device that separates either proteins or peptides by isoelectric point (pI). Sample recovery is in the solution phase. OFFGEL electrophoresis (OGE) provides the protein or peptide pI, which you can then use to validate the results of protein database searches. In the case of this sample, we know that the peptide pI ranges from 3.5 to 4.5, and we use this information during autovalidation.

This chapter assumes that the Spectrum Mill workbench has already been installed on your server, the **NCBInr** database has been downloaded and indexed, and the server is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

Exercise 1. (Optional) Transfer data files to Spectrum Mill server

In this exercise, you move Q-TOF data into the appropriate directory on the Spectrum Mill server.

NOTE The Q-TOF example data may have been installed on the server at the time the Spectrum Mill workbench was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

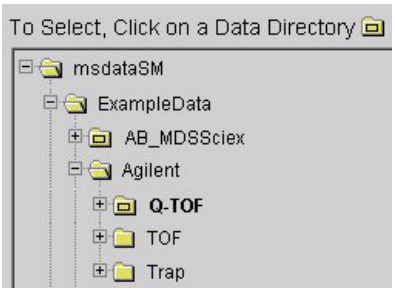
Steps	Detailed Instructions	Comments
1 Copy the example data directory, **X:\ExampleData\msdataSM\ExampleData\Agilent\Q-TOF**, from the *Spectrum Mill Example Data CD* onto your client PC.	• Copy the **Q-TOF** directory from the *Spectrum Mill Example Data CD* to any directory on your client PC.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • This exercise simulates the likely laboratory scenario of file transfer from a client or instrument PC to the server; you may transfer files directly from the CD to the server if you prefer.
2 Create the **msdataSM\ExampleData\Agilent\Q-TOF** folder on the Spectrum Mill server.	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In **msdataSM**, create **ExampleData\Agilent\Q-TOF**.	• If you don't know how to find your Spectrum Mill file system, ask the person who installed your software. • **Do not include spaces or parentheses in your directory names.** • When you process your own samples, remember to set up a separate folder for each sample. Each folder should contain all sample fractions. • You may create up to ten folders between **msdataSM** and your data files, but shorter path lengths reduce memory usage, especially for large data sets.

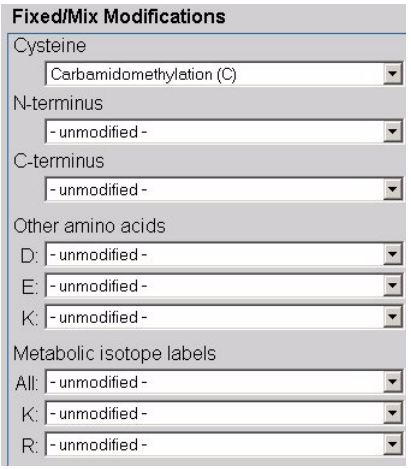
Steps	Detailed Instructions	Comments
3 Copy the example files to the new directory.	• Copy or move the contents of the **Q-TOF** directory from your client PC to the new folder on the server PC.	When you process your own data, remove any spaces or parentheses in the data file names. For example, change the file **my sample.d** to **my_sample.d**
4 Make sure you have both read and write permissions for the data folders you just created.	**a** Right-click the **Q-TOF** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked, and apply the changes to all subfolders and files. **c** Click the **Security** tab. **d** Make sure all user groups have full permissions. **e** Click **OK**.	If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors.

Exercise 2. Extract the spectra

The Spectrum Mill Data Extractor preprocesses raw data files to extract high-quality spectra for database searches. For details about the processing that the Data Extractor accomplishes, see the introductory paragraph on page 16.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link.	

Steps	Detailed Instructions	Comments
2 Select the **msdataSM\ExampleData\Agilent\Q-TOF** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **Q-TOF** to select that data file directory.  **c** Make sure that the name of the **Q-TOF** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 42 on page 120.	• Directories are identified by different types of icons: ◦ Plain folders indicate directories that do not have data files directly beneath them. ◦ Folders with rectangles indicate data directories. ◦ Folders with line spectra (bar graphs) indicate data files (which for Agilent files are also directories in themselves). • Try clicking the names of each type to see which turn **bold**, indicating that they are selectable. • The software remembers your data file selection when you go to other Spectrum Mill forms. • If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Click **OK.** The name of the modification appears as in Figure 42 on page 120.	• Carbamidomethylation is the default when you first install the Spectrum Mill workbench. • To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
4 Set remaining parameters as shown in Figure 42 on page 120.	**a** Keep the defaults whenever appropriate. This example uses all default settings, except for **Data Directory**. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. You may also need to change the time range for **Merge scans with same precursor m/z** to be compatible with your chromatographic data. **c** Click a blue section divider bar to display help for that section of the form.	• You use the **Sequence tag length** to filter out noisy spectra. This is the longest sequence of amino acids that is represented by the fragments in a spectrum. • If you set this to **>1**, you ensure that all possible good spectra are extracted. You have opportunities to set more stringent requirements later when you perform the database search.

Steps	Detailed Instructions	Comments
5 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** area to the right of the Data Extractor form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.
6 (Optional) View the **cpick_in** subdirectory to see the files Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\Agilent\Q-TOF\cpick_in**. **b** Notice the new files created there. Each one represents an extracted spectrum from your raw data file.	• File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where ◦ aaaa = first merged scan ◦ bbbb = last merged scan ◦ c = assigned precursor charge (0 means charge was ambiguous). • Because the resolution for Agilent Q-TOF data is sufficient to resolve isotopes for most peptide precursors, the great majority of extracted spectra have assigned charges.

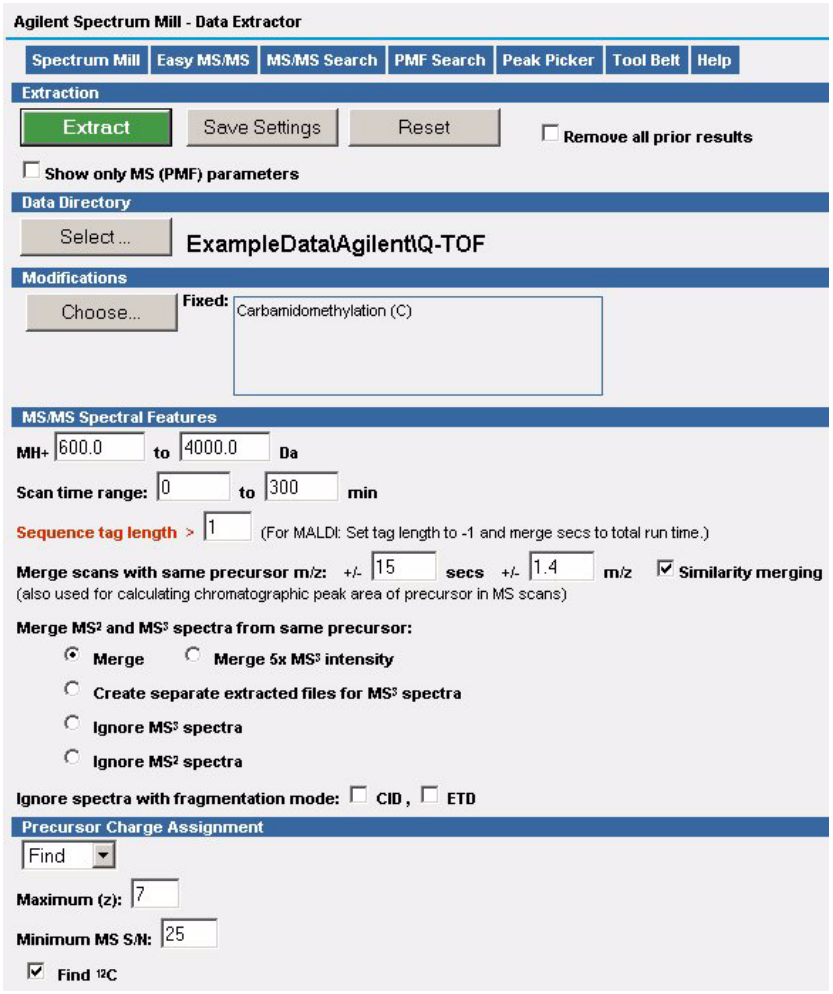
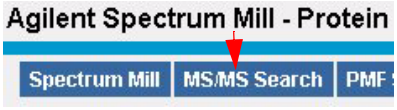


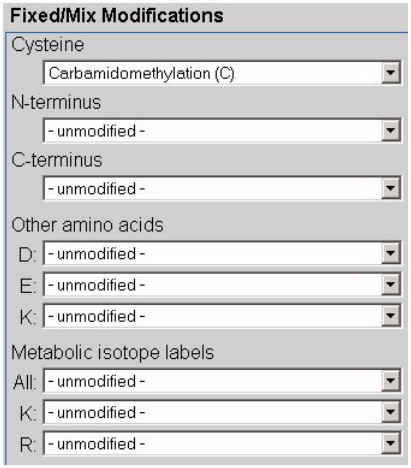
Figure 42 Settings for Data Extractor with Agilent Q-TOF data

Exercise 3. Search a database

After you have extracted your spectra, you are ready to search each spectrum against a protein or DNA database. As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal of identifying as many spectra as possible. For this initial search, you use identity mode, which means the spectra must be consistent with the database sequence. For later searches, you may use variable modifications or homology modes, which means spectra may show modifications relative to the database sequence. After you have searched for suspected modifications, you may search for unsuspected ones. See Chapter 5, "Searching for Unknown Modifications," starting on page 147.

Steps	Detailed Instructions	Comments
1 Navigate to the MS/MS Search page.	• Do one of the following: ◦ From the Protein Databases page, click the **MS/MS Search** button.  ◦ From the Spectrum Mill home page, click the link to MS/MS Search. 	Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Protein Databases page to the MS/MS Search page.
2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed the data extraction, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Select the **NCBInr.ecoli** database.	• For **Database** (under **Search Parameters**) select **NCBInr.ecoli**.	If this database is not on your server, do "Exercise 4. Create a species subset database" on page 22.

Steps	Detailed Instructions	Comments

4 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Click **OK.** The name of the modification appears as in Figure 43 on page 124.	• The modification is likely already set.
		• Three types of modifications are available for MS/MS Search:
		  ◦ The **fixed** modifications are assumed to apply universally and are searched in a single search cycle.
		  ◦ The **mix** modifications trigger cyclic MS/MS searches, where a different form of the modification is searched in each cycle.
		  ◦ The **variable** modifications allow for both modified and unmodified forms within a peptide. All variable modifications are searched in each cycle.
		• For more information about choosing modifications, see the online help.
		• Your system administrator can configure custom modifications.

5 Set up the **Matching Tolerances**.	**a** Change the **Instrument** setting to **Agilent ESI Q-TOF**. **b** Leave the other settings at default values.	

Steps	Detailed Instructions	Comments
6 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Dynamic peak thresholding**. **c** Verify that the **Search mode** is set to **Identity**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.
7 Set other parameters as shown in Figure 43 on page 124.	**a** Keep the default settings. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	Changes from defaults are highlighted in yellow.
8 Start the search.	**a** Click the **Start Search** button. **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • Search time varies depending on the size of the database searched. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

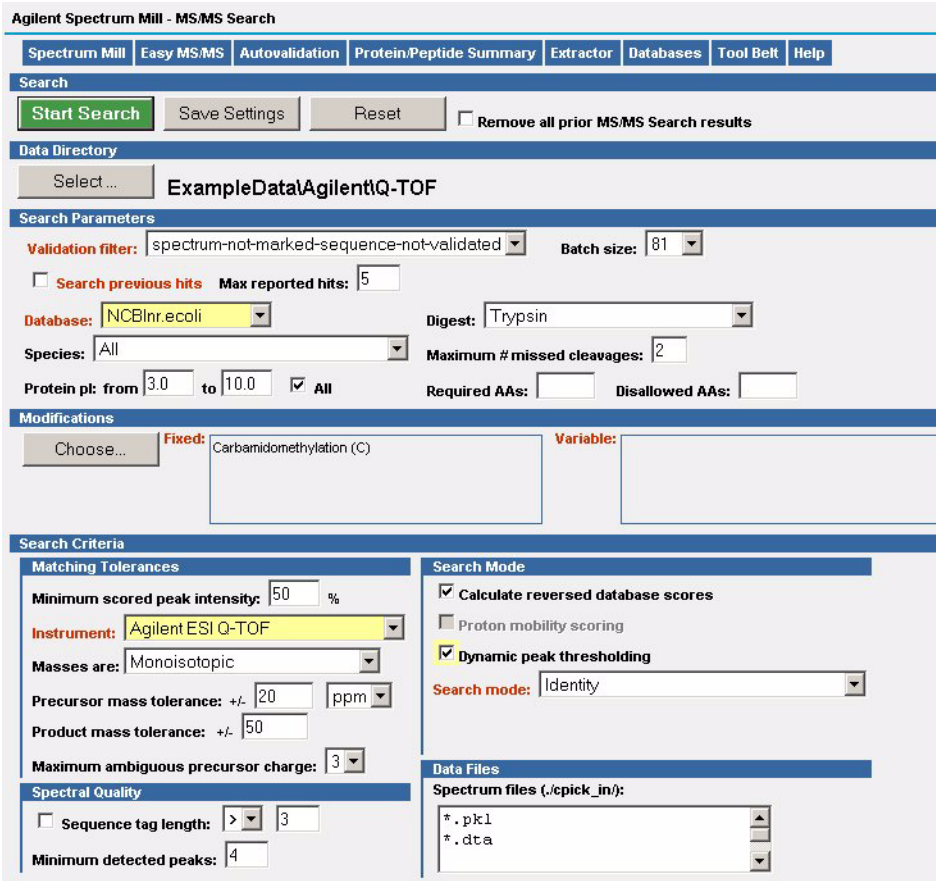


Figure 43 MS/MS Search settings

Exercise 4. Autovalidate the hiqhest-quality results

After you have completed a database search, you validate the good results. Validation means that you accept that the matches are correct.

The Spectrum Mill workbench provides a means to segregate search results that contain a valid interpretation of an MS/MS spectrum from those that do not. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

The Spectrum Mill workbench provides two ways to validate results. One way uses the Autovalidation page, and is totally automated. You use this method only to validate the highest-scoring results — those that do not require manual review. The other method of validating uses the Protein/Peptide Summary page for manual review and validation. This exercise describes autovalidation.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 44 on page 127.

2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed the data extraction and MS/MS Search for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	

Steps	Detailed Instructions	Comments
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Mark the check box for **Filter by Peptide pI**, and type `3.5` for **Low** and `4.5` for **High**. (This is the pI range of the sample fraction from the Agilent OFFGEL Fractionator.) **c** Keep the defaults for the other settings, as shown in Figure 44 on page 127. **d** Click the **Validate Files** button. **e** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein. • Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20. • To contribute to the protein score, the peptide pI must also fall into the 3.5 to 4.5 range. • If you wish to use the pI filter for modified peptides, ask your server administrator to first verify that the pK of the modified amino acid is specified in **smconfig.std.xml** or **smconfig.custom.xml**.
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Check that **Filter by Peptide pI** matches what you see in Figure 45 on page 128. **c** Keep the defaults for the other settings, as shown in Figure 45. **d** Click the **Validate Files** button. **e** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by peptide. Even if it finds only a single peptide for a protein, it validates the corresponding search results provided that the peptide score is high enough. • Using the default scoring, individual peptides must have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein. • Again, the peptide pI must fall into the 3.5 to 4.5 range or the peptide is not validated.

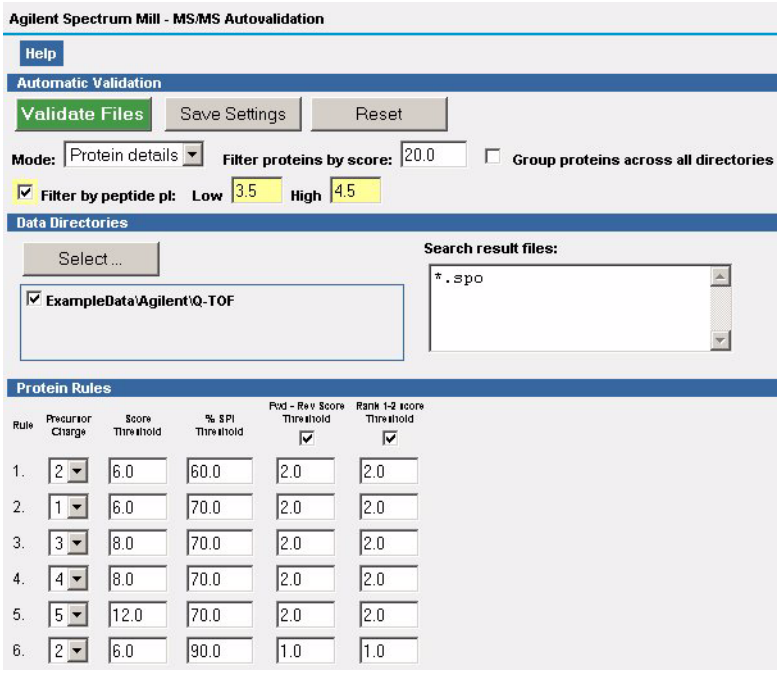


Figure 44 Autovalidation form, Protein details mode

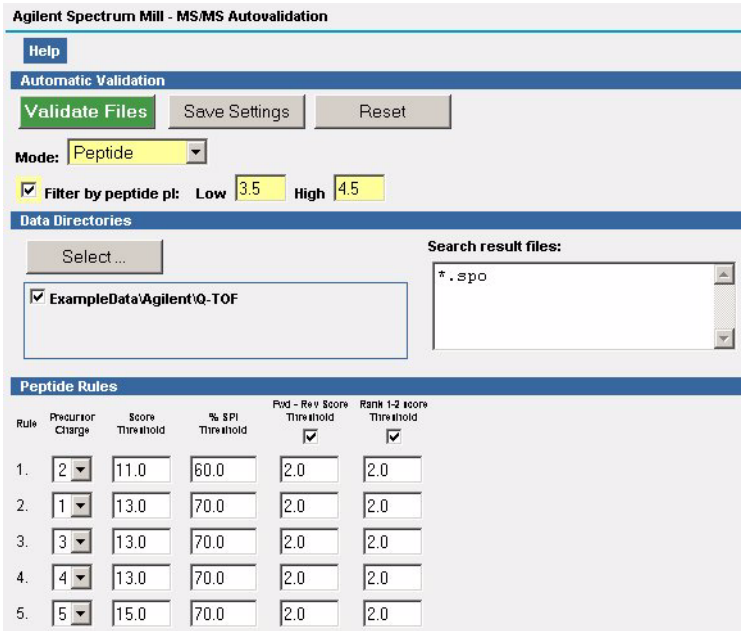


Figure 45 Autovalidation form, Peptide mode

Exercise 5. (Optional) Examine the validated results

If this is the first time you have performed an autovalidation, you may want to review the validated data to convince yourself that autovalidation really works.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page, if you are not already there.	• From the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other Spectrum Mill pages.
2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed autovalidation for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Set the **Validation preset** to **none**.		

Steps	Detailed Instructions	Comments
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	• By setting **Score** and **% SPI** to zero, you ensure that all valid data is selected, regardless of score.
8 Set the **Review Fields**.	**a** Mark the check box for **Rank 1-2 score**. **b** Mark the check box for **Peptide pI**. **c** For others, keep the default settings. **d** Check to see that your form looks like that in Figure 46 on page 131. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		

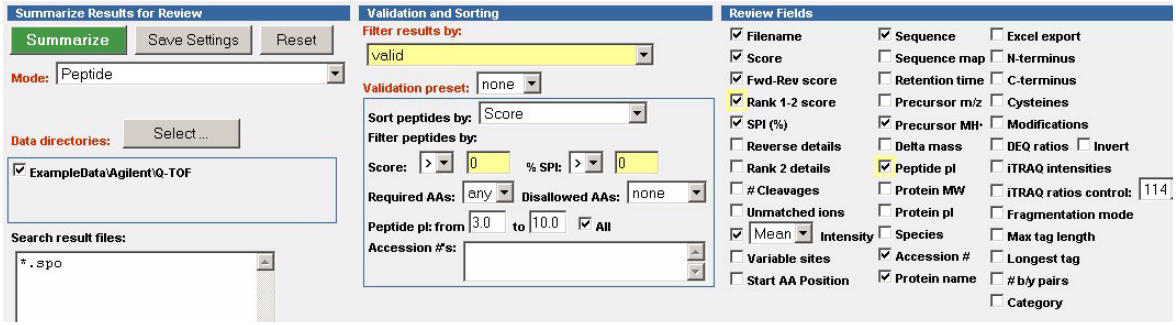


Figure 46 Settings to check autovalidation results

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Q-TOF
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 311 **Hits:** 311

#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	Spectrum Intensity	Sequence	MH+ Matched (Da)	Peptide pI	Accession #	Protein Name
1	ec-oge-2.1605.1613.2	2	26.23	26.23	26.23	99.8	1.17e+006	(R)VITASDSSGTVVDESGFTK(E)	1899.918	4.03	16129715	glutamate dehydrogenase, NADP-specific
2	ec-oge-2.2234.2237.2	2	25.99	25.99	25.99	99.8	3.15e+005	(K)FIEQDPEGQYGLEAAFR(S)	1969.929	4.00	26250758	DNA-directed RNA polymerase beta subunit
3	ec-oge-2.2534.2542.2	2	25.12	25.12	25.12	97.1	9.33e+005	(R)IQQQQDIDNGTLPDFISETASIR(D)	2589.279	3.84	16131840	malate synthase A
4	ec-oge-2.2658.2666.2	2	25.00	25.00	25.00	100.0	5.37e+005	(K)DFEAQLASTETQVGNELAPLK(G)	2261.130	4.00	7449165	adhesin homolog yebL precursor
5	ec-oge-2.1922.1929.2	2	24.99	24.99	24.99	99.9	1.15e+006	(R)LYTTNADGELITIDTADNK(I)	2068.008	3.84	15801691	putative receptor
6	ec-oge-2.2008.2016.2	2	24.88	24.88	24.88	98.3	2.41e+006	(K)VEVETPEENTGDVIGDLSR(R)	2058.983	3.77	15803853	elongation factor EF-2
7	ec-oge-2.2210.2217.2	2	24.61	24.61	24.61	96.0	5.87e+005	(K)FSLPDQDGEQVNLTDFQGQR(V)	2294.068	3.84	15803003	bacterioferritin comigratory protein
8	ec-oge-2.2394.2402.2	2	24.50	13.36	24.50	100.0	6.00e+005	(R)IEPSLEAAFVDYGAER(H)	1766.860	4.00	42773	unnamed protein product
9	ec-oge-2.2688.2696.2	2	24.24	24.24	24.24	99.9	8.45e+005	(K)LSDQEIEQTLQAFEAR(V)	1877.924	4.00	15803860	FKBP-type peptidyl-prolyl cis-trans isomerase (rotam
10	ec-oge-2.2753.2760.2	2	23.86	23.86	23.86	95.6	8.01e+005	(R)DVPTNEGVLGEIALSSLPR(I)	1967.044	4.14	16131102	glutamate synthase, large subunit
11	ec-oge-2.2581.2588.2	2	23.75	23.75	23.75	99.9	5.15e+005	(R)EIPSDIVYQDDLVTAFR(D)	1980.991	3.84	15801220	hypothetical protein Z1742
12	ec-oge-2.1702.1708.2	2	23.49	23.49	23.49	99.8	6.88e+005	(R)ALSNPDLYEGDGELR(V)	1648.781	3.91	15800202	heat shock protein 90
13	ec-oge-2.2761.2768.2	2	23.49	23.49	23.49	94.8	7.87e+005	(K)GFSGEDATPALEGADVVLISAGVAR(K)	2402.220	3.91	15803770	malate dehydrogenase
14	ec-oge-2.2376.2384.2	2	23.33	23.33	23.33	98.0	1.18e+006	(R)LAFDEFDELAGDR(A)	1497.686	3.77	15799867	acetyl-CoA carboxylase alpha subunit
15	ec-oge-2.2260.2268.2	2	23.32	23.32	23.32	97.0	5.42e+005	(R)LDFCNLDVNDTAAFSR(L)	1857.844	3.93	16129805	glucose-6-phosphate dehydrogenase
16	ec-oge-2.2924.2932.2	2	23.28	23.28	23.28	95.8	9.07e+005	(K)ANDAAGDGTTTATVLAQAIITEGLK(A)	2402.241	4.03	15804735	chaperonin GroEL
17	ec-oge-2.1838.1844.2	2	23.27	23.27	23.27	95.7	5.71e+005	(R)YDQDPANQGIVDALK(A)	1646.802	3.93	16131330	leucine transporter subunit
18	ec-oge-2.2654.2660.2	2	23.04	23.04	23.04	97.3	2.43e+006	(R)DATGIDPVSLIAFDK(V)	1561.811	3.93	15803846	50S ribosomal protein L4
19	ec-oge-2.2438.2446.2	2	22.90	22.90	22.90	93.9	2.86e+005	(R)IDAAFQDEVAASEGFLK(Q)	1810.886	3.91	15802856	histidine-binding periplasmic protein of high-affinity hi
20	ec-oge-2.1357.1363.2	2	22.75	22.75	22.75	99.9	1.48e+006	(K)ISADQVDQEVER(F)	1388.665	3.91	15802949	PEP-protein phosphotransferase system enzyme I
21	ec-oge-2.1977.1990.2	2	22.66	22.66	22.66	95.4	5.68e+005	(K)DTTTIIDGVGEEAAIQGR(V)	1845.919	3.91	15804735	chaperonin GroEL
22	ec-oge-2.2046.2052.2	2	22.64	22.64	22.64	98.3	3.89e+006	(R)VDDLAVDLVER(F)	1243.653	3.84	16128870	pyruvate formate lyase I
23	ec-oge-2.2397.2406.2	2	22.53	22.53	22.53	97.5	7.61e+005	(K)ISDEQLDQAIANIAK(Q)	1628.849	4.03	15799738	survival protein
24	ec-oge-2.2926.2936.2	2	22.48	22.48	22.48	99.8	7.93e+005	(K)LSDAEVDELFALVK(A)	1548.816	3.91	16128065	3-isopropylmalate isomerase subunit
25	ec-oge-2.2202.2208.2	2	22.47	22.47	18.14	99.0	2.40e+006	(R)LVPFEDLPDEEVGSR(E)	1701.833	3.83	16131102	glutamate synthase, large subunit
26	ec-oge-2.2160.2168.2	2	22.44	22.44	22.44	92.8	1.25e+006	(K)ADGINPEELLGNSSAAAPR(A)	1881.930	4.14	15803187	DNA-binding protein; H-NS-like protein; chaperone ac

Figure 47 Summary of valid results from autovalidation (partial list)

Steps	Detailed Instructions	Comments
10 Examine the summary report.	**a** Check that your results are similar to those in Figure 47 on page 131. Since the public databases are constantly updated, your results may vary slightly. **b** Verify that the peptide pI values fall between 3.5 and 4.5. **c** Become familiar with the following headers: • **#** - table entry number • **Filename** - the name of the spectral file • **z** - peptide charge • **Score** - peptide database search score • **Fwd-Rev Score** - difference between scores for top hits from forward and reversed database searches • **Rank1-Rank2 Score** - difference between rank 1 and rank 2 database search scores • **SPI (%)** - % scored peak intensity, which is the percentage of the ion current in the extracted MS/MS spectrum that is explained by the top database hit • **Spectrum Intensity** - intensity calculated from extracted ion chromatogram of precursor ion • **Sequence** - sequence from top database hit • **MH^+ Matched (Da)** - measured precursor ion MH^+ • **Peptide pI** - the calculated pI of the peptide • **Accession #** - database accession number for the top hit • **Protein Name** - name of the top database hit	• File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where ◦ aaaa = first merged scan ◦ bbbb = last merged scan ◦ c = assigned precursor charge (0 means charge was ambiguous). • Because the resolution for Agilent Q-TOF data is sufficient to resolve isotopes for most peptide precursors, the great majority of extracted spectra have assigned charges.

Steps	Detailed Instructions	Comments
11 To examine these results in more detail, go to "Exercise 10. Examine results with the Spectrum Viewer" on page 44 of Chapter 1.		

Exercise 6. Display medium-quality results in preparation for manual validation

Now that you have automatically validated the best results with autovalidation, you prepare to validate medium-quality results through a process of manual inspection and validation. Manual validation helps you to maximize the number of proteins you identify in your samples, or to increase the sequence coverage of the identified proteins, but it does require that you devote some time. When you process your own samples, you may skip this step if automatic validation provides sufficient protein identifications for the purposes of your experiment.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button near the top of the form to display the settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other pages.
2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed autovalidation or results review for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Steps	Detailed Instructions	Comments
4 For **Filter results by**, reset to the default setting of **sequence-not-validated**.		• **Filter results by** selects from your data only the results that match your setting. • In this case, it selects only the results that have not yet been validated during autovalidation.
5 Set the **Validation preset** to **reset**.		• You always want to set the **Validation preset** to the setting most likely to be accurate for the majority of the data. This minimizes manual changes later. • Since you will set score thresholds to select the medium-quality results, the **reset** setting is best for this round of processing. That way you manually change the **Validation category** setting to **valid** (**V**) for those matches you accept as being correct.
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)

Steps	Detailed Instructions	Comments
7 Set the **Review Fields**.	**a** Mark the check box for **Rank 1-2 score**. **b** Mark the check box for **Peptide pI**. **c** Mark the check boxes for **# Cleavages** and **Unmatched ions**. **d** For other **Review Fields**, keep the default settings.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. Differences of at least 1 and preferably 2 can indicate a good match, especially for the lower-scoring hits. • The **# Cleavages** is the number of cleavages of the amino acid backbone that are represented in the spectrum for the assigned sequence. For CID spectra, these cleavages include the b- and y-fragments, and for ETD spectra, they include the c- and z-fragments. Other ions, such as neutral losses from b and y ions, are not included. • When you use **Dynamic peak thresholding** in the search it is helpful to mark the check box for **Unmatched ions**. These results tell you both the number of unmatched ions and the number of ions that were included in the search score that you see in the report. If the latter number is less than 15, check that the peptide is either short or in low abundance. If not, be suspicious of the results. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.

Steps	Detailed Instructions	Comments
8 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 5**, to capture all results that may possibly be good. **b** For **% SPI**, keep the default of **> 60**. Lower values suggest poor results. **c** For **Peptide pI**, type a range from **3.5** to **4.5** and *clear* the check box for **All**. **d** Check to see that your form looks like that in Figure 48. Note that settings that have been changed from defaults are highlighted in yellow.	• These parameters select from the data set only the members that exhibit the specified score and SPI criteria. • These settings select results of medium quality or better. • The peptide pI is known from the Agilent OFFGEL Fractionator and serves as an additional validation criterion. • Remember that the **Filter results by** setting prevents display of the highest-quality results (the ones that have already been validated).
9 Click the **Summarize** button.		

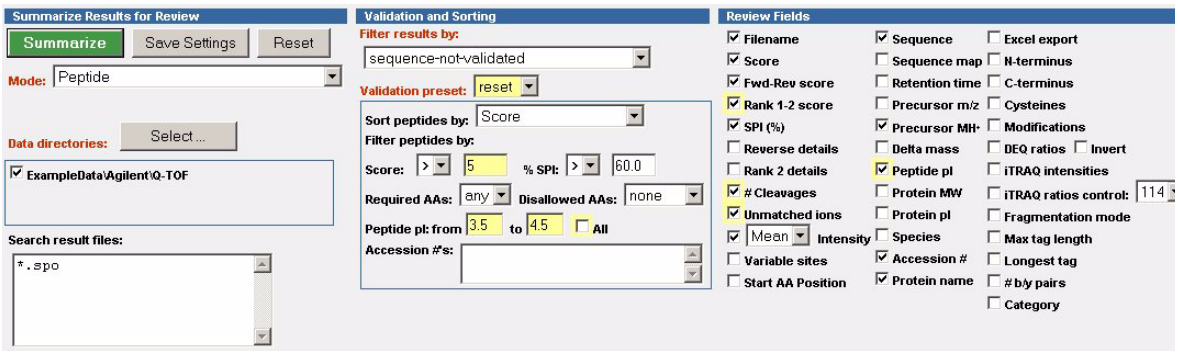
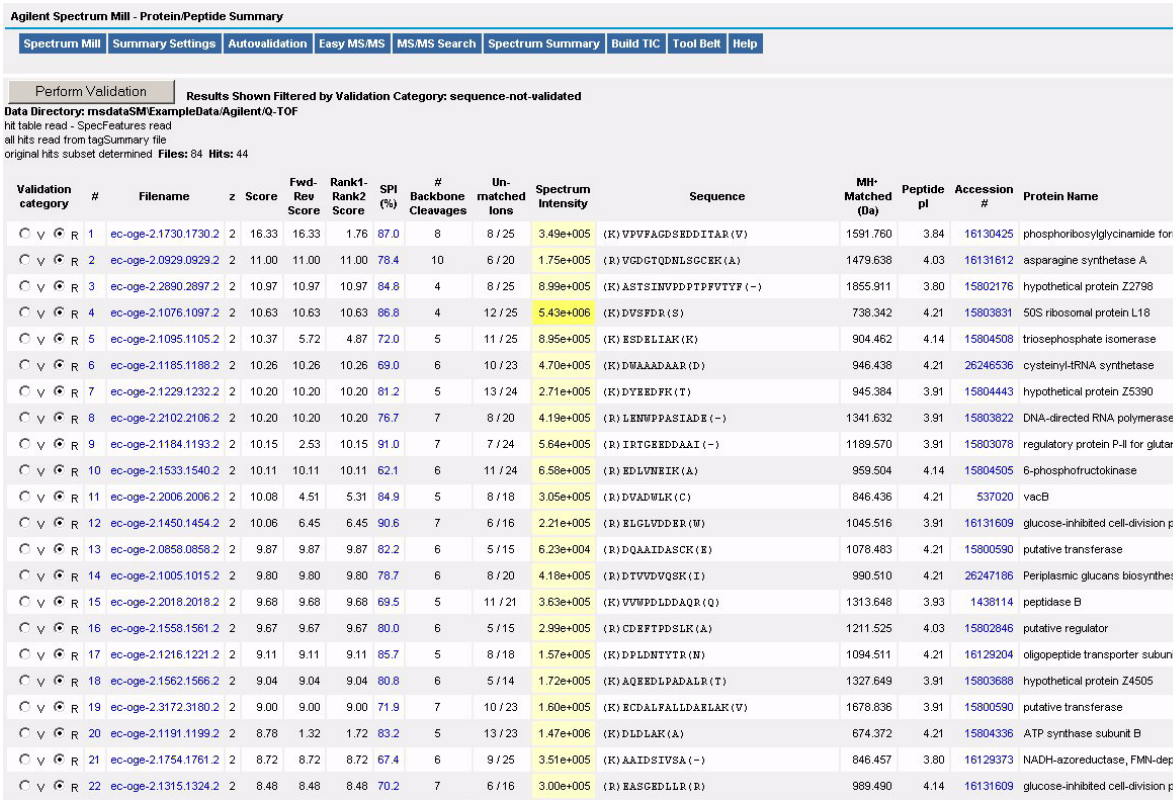


Figure 48 Protein/Peptide Summary settings to display medium-quality results

Steps	Detailed Instructions	Comments
10 Examine the summary report.	• Check to see that your results are similar to those in Figure 49. Note that since the public databases are constantly updated, your results may vary slightly.	• File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where ◦ aaaa = first merged scan ◦ bbbb = last merged scan ◦ c = assigned precursor charge (0 means charge was ambiguous). • Because the resolution for Agilent Q-TOF data is sufficient to resolve isotopes for most peptide precursors, the great majority of extracted spectra have assigned charges.
11 To examine these results in more detail, go to "Exercise 7. Manually review and validate results" on page 140.		
12 (Optional) Vary input parameters and investigate the results.	**a** To change parameters and create a new summary table, click the **Summary Settings** button near the top of the form to start over. **b** Before proceeding to the next exercise, be sure to go back to the settings you had in step 8 of this exercise.	



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Perform Validation — Results Shown Filtered by Validation Category: sequence-not-validated

Data Directory: msdataSM\ExampleData/Agilent/Q-TOF
hit table read - SpecFeatures read
all hits read from tagSummary file
original hits subset determined **Files:** 84 **Hits:** 44

Validation category	#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	# Backbone Cleavages	Un-matched Ions	Spectrum Intensity	Sequence	MH+ Matched (Da)	Peptide pI	Accession #	Protein Name
V R	1	ec-oge-2.1730.1730.2	2	16.33	16.33	1.76	87.0	8	8 / 25	3.49e+005	(K)VPVFAGDSEDDITAR(V)	1591.760	3.84	16130425	phosphoribosylglycinamide for
V R	2	ec-oge-2.0929.0929.2	2	11.00	11.00	11.00	78.4	10	6 / 20	1.75e+005	(R)VGDGTQDNLSGCEK(A)	1479.638	4.03	16131612	asparagine synthetase A
V R	3	ec-oge-2.2890.2897.2	2	10.97	10.97	10.97	84.8	4	8 / 25	8.99e+005	(K)ASTSINVPDPTPFVTYF(-)	1855.911	3.80	15802176	hypothetical protein Z2798
V R	4	ec-oge-2.1076.1097.2	2	10.63	10.63	10.63	86.8	4	12 / 25	5.43e+006	(K)DVSFDR(S)	738.342	4.21	15803831	50S ribosomal protein L18
V R	5	ec-oge-2.1095.1105.2	2	10.37	5.72	4.87	72.0	5	11 / 25	8.95e+005	(K)ESDELIAK(K)	904.462	4.14	15804508	triosephosphate isomerase
V R	6	ec-oge-2.1185.1188.2	2	10.26	10.26	10.26	69.0	6	10 / 23	4.70e+005	(K)DWAAADAAR(D)	946.438	4.21	26246536	cysteinyl-tRNA synthetase
V R	7	ec-oge-2.1229.1232.2	2	10.20	10.20	10.20	81.2	5	13 / 24	2.71e+005	(K)DYEEDFK(T)	945.384	3.91	15804443	hypothetical protein Z5390
V R	8	ec-oge-2.2102.2106.2	2	10.20	10.20	10.20	76.7	7	8 / 20	4.19e+005	(R)LENWPPASIADE(-)	1341.632	3.91	15803822	DNA-directed RNA polymerase
V R	9	ec-oge-2.1184.1193.2	2	10.15	2.53	10.15	91.0	7	7 / 24	5.64e+005	(R)IRTGEEDDAAI(-)	1189.570	3.91	15803078	regulatory protein P-II for glutar
V R	10	ec-oge-2.1533.1540.2	2	10.11	10.11	10.11	62.1	6	11 / 24	6.58e+005	(R)EDLVNEIK(A)	959.504	4.14	15804505	6-phosphofructokinase
V R	11	ec-oge-2.2006.2006.2	2	10.08	4.51	5.31	84.9	5	8 / 18	3.05e+005	(R)DVADWLK(C)	846.436	4.21	537020	vacB
V R	12	ec-oge-2.1450.1454.2	2	10.06	6.45	6.45	90.6	7	6 / 16	2.21e+005	(R)ELGLVDDER(W)	1045.516	3.91	16131609	glucose-inhibited cell-division p
V R	13	ec-oge-2.0858.0858.2	2	9.87	9.87	9.87	82.2	6	5 / 15	6.23e+004	(R)DQAAIDASCK(E)	1078.483	4.21	15800590	putative transferase
V R	14	ec-oge-2.1005.1015.2	2	9.80	9.80	9.80	78.7	6	8 / 20	4.18e+005	(R)DTVVDVQSK(I)	990.510	4.21	26247186	Periplasmic glucans biosynthes
V R	15	ec-oge-2.2018.2018.2	2	9.68	9.68	9.68	69.5	5	11 / 21	3.63e+005	(K)VVWPDLDDAQR(Q)	1313.648	3.93	1438114	peptidase B
V R	16	ec-oge-2.1558.1561.2	2	9.67	9.67	9.67	80.0	6	5 / 15	2.99e+005	(R)CDEFTPDSLK(A)	1211.525	4.03	15802846	putative regulator
V R	17	ec-oge-2.1216.1221.2	2	9.11	9.11	9.11	85.7	5	8 / 18	1.57e+005	(K)DPLDNTYTR(N)	1094.511	4.21	16129204	oligopeptide transporter subun
V R	18	ec-oge-2.1562.1566.2	2	9.04	9.04	9.04	80.8	6	5 / 14	1.72e+005	(K)AQEEDLPADALR(T)	1327.649	3.91	15803688	hypothetical protein Z4505
V R	19	ec-oge-2.3172.3180.2	2	9.00	9.00	9.00	71.9	7	10 / 23	1.60e+005	(K)ECDALFALLDAELAK(V)	1678.836	3.91	15800590	putative transferase
V R	20	ec-oge-2.1191.1199.2	2	8.78	1.32	1.72	83.2	5	13 / 23	1.47e+006	(K)DLDLAK(A)	674.372	4.21	15804336	ATP synthase subunit B
V R	21	ec-oge-2.1754.1761.2	2	8.72	8.72	8.72	67.4	6	9 / 25	3.51e+005	(K)AAIDSIVSA(-)	846.457	3.80	16129373	NADH-azoreductase, FMN-dep
V R	22	ec-oge-2.1315.1324.2	2	8.48	8.48	8.48	70.2	7	6 / 16	3.00e+005	(R)EASGEDLLR(R)	989.490	4.14	16131609	glucose-inhibited cell-division p

Figure 49 Summary of medium-quality results (partial list)

Exercise 7. Manually review and validate results

After you display medium-quality results, you manually review them. In this exercise, you examine the MS/MS search results to determine how well they fit the experimental spectra. As you examine the results, you designate good results as **valid**.

Steps	Detailed Instructions	Comments
1 Display the first spectrum.	**a** In the summary report, click the number **1** under the **#** header. **b** Check the bottom of the browser window, where you see the first spectrum displayed in the Spectrum Viewer. **c** Check that your results are similar to those in Figure 50 on page 141. Since the public databases are constantly updated, your results may vary. **d** Now turn your attention again to the summary report. **e** Click the first number under the **SPI** header. **f** Check that you see the same spectrum displayed in the Spectrum Viewer as you did in substep **b**, only this time the Spectrum Viewer appears in a new window.	The Spectrum Viewer shows your extracted MS/MS spectrum. This spectrum is annotated with theoretical fragments from the peptide identified as the best database match.
2 Use the Spectrum Viewer to examine details of the MS/MS search interpretation and its correlation with the spectrum.	**a** Learn to use the Spectrum Viewer. Follow the steps in "Exercise 10. Examine results with the Spectrum Viewer" on page 44 of Chapter 1. **b** Use the Spectrum Viewer to evaluate the search results.	In the spectrum, unmatched ions are shown in black, while matched ions are displayed in red and blue. Precursor ions, neutral losses from precursor ions, and immonium ions are shown in green.

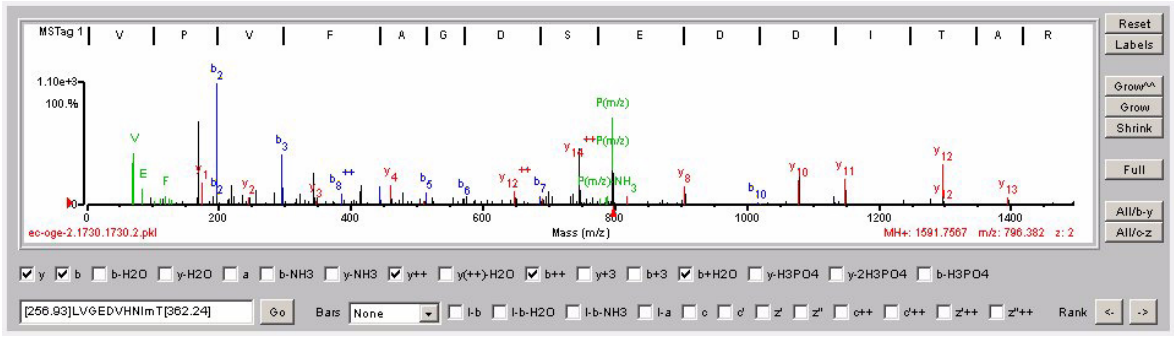


Figure 50 Spectrum Viewer

Steps	Detailed Instructions	Comments
3 Decide whether to mark the database search results for the first spectrum as **valid**.	**a** For purposes of this exercise, you decide that the protein sequence is a good match for the spectrum. **b** Under the **Validation category** heading, click the **V** to set the validation preset to **valid**.	
4 Examine additional spectra.	**a** Repeat step 2 (and optionally step 3) for as many spectra as you wish to examine. **b** Do not change any more of the validation presets under the **Validation category** heading. If you do, your results will not match in later exercises.	• When you process your own data, you will likely review all the spectra. • When you process your own data, you will change the **Validation category** presets from **reset** to **valid** (**R** to **V**) as you encounter good results.
5 Click the **Perform Validation** button.	• Check that you see a validation summary that lists all the hits and spectra that have been validated thus far (from both autovalidation and manual validation).	This makes your assessments a permanent part of the validation record.

Exercise 8. Summarize valid results

Now you summarize all the results you have validated thus far. These include results validated both automatically and manually.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button near the top of the page to display settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	
2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed results review for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Set display **Mode** to **Protein Summary**.		• When you process your own samples, you may choose another display mode, depending on whether you want to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set the **Filter results by** to **valid**.		• **Filter results by** selects from your data only the results that match your setting. • In this case, it selects only the valid results.

Steps	Detailed Instructions	Comments
5 For **Sort proteins by**, keep the default of **Score**.		• The **Sort** parameters determine how the data are sorted in the results display. • When you process your own data, select the setting that is most helpful to you.
6 Set the **score** and **% SPI** filters to **> 0**.	**a** Set **Filter by protein score** to **> 0**. **b** For **Filter peptides by**, set both **Score** and **% SPI** to **> 0**.	When you set these filters to **> 0**, you are sure to display all the valid results.
7 Set the **Review Fields**.	**a** Mark the **Protein MW** check box. **b** Keep the other default settings. **c** Check to see that your form looks like that in Figure 51. Note that settings that are changed from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
8 Click the **Summarize** button.		

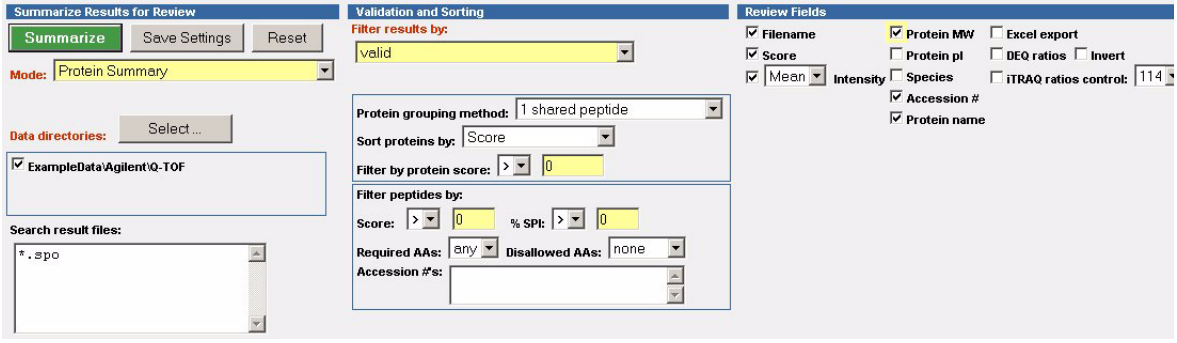


Figure 51 Protein/Peptide Summary settings to display a final results summary

Steps	Detailed Instructions	Comments
9 Examine the overall summary report.	• Check that your results are similar to those in Figure 52 on page 145.	• Since the public databases are constantly updated, your results may vary, • Your results also vary depending on which peptides you manually validated.
10 Note the colored cells in the summary report.		• The color code in the **Distinct Peptides** column indicates the relative number of peptides detected for each protein. Darker colors represent larger numbers of peptides. • The color code in the **Mean Peptide Spectral Intensity** column indicates relative protein concentrations. Darker colors represent higher concentrations. The number in each cell is an average of the intensities for all the peptides assigned to that protein. The peptide intensities are calculated from extracted ion chromatograms from the precursor ions. • For more accurate quantitation, under **Review Fields**, for **Intensity**, select **Total** rather than **Mean**. • The intensity results are sufficient for studies where you are interested in differences of two-fold or more.
11 (Optional) Vary the summary settings (particularly the **Mode** setting) and investigate the results.		See the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Q-TOF
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 312 **Hits:** 1101
beginning to assemble proteins .... proteins assembled 0.1109 sec
proteins filtered by unique peptides 0.227067 sec
proteins filtered by score
calculated protein coverage maps 1.629192 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 1.967531 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 217 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	Protein MW (Da)	Database Accession #	Protein Name
1	5	5	85.33	12	1.10e+006	58957.1	26246004	2-isopropylmalate synthase
2	4	4	77.97	4	1.40e+006	167284.4	26249799	Glutamate synthase [NADPH] large chain precursor
3	4	4	70.47	6	3.84e+005	87435.6	16129658	phosphoenolpyruvate synthase
4	3	3	60.94	6	5.43e+005	71420.9	15800202	heat shock protein 90
5	3	3	59.75	10	7.18e+005	57357.2	45686198	GroEL
6	3	3	58.98	13	2.33e+006	46446.2	563868	enolase; phosphopyruvate hydratase
7	3	3	54.71	9	1.52e+007	44850.6	26249935	elongation factor Tu
8	4	3	54.30	4	4.16e+005	118344.4	26247224	Ribonuclease E
9	3	3	52.78	6	1.61e+006	59597.8	26250514	ketol-acid reductoisomerase
10	4	4	52.32	16	1.03e+007	20761.5	15800320	alkyl hydroperoxide reductase, C22 subunit; detoxification
11	4	3	51.90	5	3.91e+005	99669.0	15799798	pyruvate dehydrogenase (decarboxylase component)
12	3	3	51.73	4	7.65e+005	96405.4	15803211	alanyl-tRNA synthetase
13	4	3	49.71	10	6.72e+006	42934.7	33383669	isocitrate dehydrogenase
14	3	3	49.67	6	8.58e+005	77095.7	26248538	Colicin I receptor precursor
15	3	3	45.95	4	9.63e+005	72532.3	15801890	exoribonuclease II

Figure 52 Final results summary (partial listing), in Protein Summary mode

5
Searching for Unknown Modifications

Exercise 1. Create a list of identified proteins 149
Exercise 2. Search for unknown modifications 151
Exercise 3. Autovalidate the highest-quality results 156
Exercise 4. Examine the list of validated peptides 158
Exercise 5. Examine details of search results to determine modification sites 163

This chapter shows how to search MS/MS spectra to identify unsuspected modifications. If you expect that certain modifications are present in your sample, then you search specifically for those. (See "Exercise 2. Search in variable modifications mode" on page 64 of Chapter 2.) To search for modifications that you do not suspect, you conduct a "mass gap" search. This search is named "mass gap" because the modification produces a gap (a mass shift) between the masses in the experimental MS/MS spectrum and the spectrum that would result from the unmodified sequence in the database. It is best to perform variable modifications searches prior to a mass gap search because variable modifications searches are more specific and reduce the chance of false positive identifications.

This chapter uses the Agilent Q-TOF example data to illustrate the mass gap search. You can conduct a mass gap search with any MS/MS data, but the chance of a false positive result diminishes greatly when MS/MS spectra exhibit greater mass accuracy. As with the variable modifications described in Chapter 2, it is best to conduct a mass gap search against a file of previously validated hits. That way, you decrease search time and minimize the chance of random matches.

After you conduct a mass gap search, you autovalidate the results. The Spectrum Mill workbench adds the newly validated results to those you validated previously.

Before you begin the exercises in this chapter, please complete the exercises in Chapter 4.

Exercise 1. Create a list of identified proteins

You can perform a mass gap search against a full database, a species-specific subset database, or a list of proteins that you have already identified in your sample. In this series of exercises, you search against the latter because the modified peptides you find with a mass gap search are often derived from the proteins you already identified in identity mode. The mass gap search is much faster and less likely to lead to false positive identifications when you search this smaller list.

This exercise shows how to create a file of previously validated hits, which is a small database of proteins that you have already identified in your sample. In this exercise, you create the file using the Tool Belt page. Alternatively, you can create the file automatically from the MS/MS Search page. When you mark the check box to **Search previous hits** in MS/MS Search, if you have not created a saved results file, then the search page automatically creates one for you. (On the form, be sure to select the database that you searched previously.)

Steps	Detailed Instructions	Comments
1 Navigate to the Tool Belt page.	• From the Spectrum Mill home page, click the link to the Tool Belt page.	There are other ways to navigate to the Tool Belt page. Check the top of your current page for a **Tool Belt** button.
2 Click the **Create saved results file** option near the top of the form.		• You will create a file of valid hits from the search results you validated in Chapter 4. • You will later search this protein list to identify spectra that show modifications.
3 Fill in the form as shown in Figure 53.	**a** Select the Q-TOF data directory that you previously searched (**ExampleData\Agilent\Q-TOF**). **b** Select the database you previously searched (**NCBInr.ecoli**).	
4 Click the **Create File** button.		

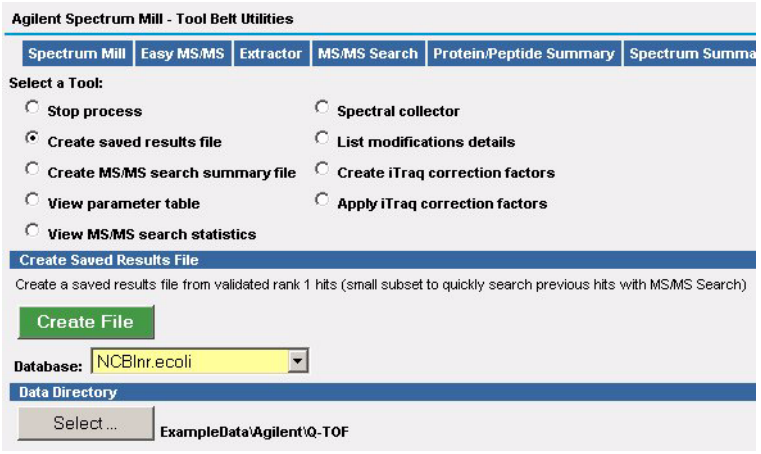


Figure 53 Create file of previously validated proteins

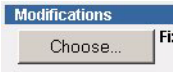
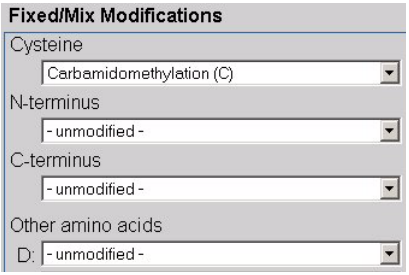
Exercise 2. Search for unknown modifications

In this exercise, you conduct a mass gap search against database hits that you validated previously.

The MS/MS Search software provides identity, variable modifications, and homology search modes. In identity mode, you search for exact matches in the selected database. In variable modifications mode, you search for post-translational and chemical modifications that occur at a fraction of the available sites. In homology modes, you search for single amino acid substitutions relative to the database (as well as any variable modifications you have selected).

When you select one of the homology modes and click the option for **Unassigned single mass gap**, the search looks for an unexpected modification (a mass gap). Note that you cannot simultaneously search variable modifications in combination with the unassigned single mass gap.

Steps	Detailed Instructions	Comments
1 Navigate to the MS/MS Search page.	• From the Tool Belt page or the Spectrum Mill home page, click the link to MS/MS Search.	There are a number of ways to navigate to the MS/MS Search page.
2 Check that your **Data Directory** is set to **ExampleData\Agilent\ Q-TOF**.	• If you have just been working with the Agilent Q-TOF data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Set **Search Parameters**.	**a** For **Database**, select **NCBInr.ecoli**, as before. **b** Mark the check box next to **Search previous hits.**	This searches the file of previous hits you created for this data directory. (See "Exercise 1. Create a list of identified proteins" on page 149.)

Steps	Detailed Instructions	Comments
4 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form.  **b** Under the **Cysteine** heading, select **carbamidomethylation**.  **c** Click **OK.** The names of the modifications appear as in Figure 54 on page 155.	• This is the default modification. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
5 Set up the **Matching Tolerances**.	**a** Change the **Instrument** setting to **Agilent ESI Q-TOF**. **b** Leave the other settings at default values.	

Steps	Detailed Instructions	Comments
6 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Dynamic peak thresholding**. **c** Change the **Search mode** to **Homology - All mutations**. **d** Click the option for **Unassigned single mass gap**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set. • You could instead set the **Search mode** to **Homology - Single base pair mutations**. The results would be the same.
7 Set the other parameters as shown in Figure 54 on page 155.	**a** Keep the defaults whenever appropriate. Changes from defaults are highlighted in yellow. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	• You do not mark the check box to **Remove all prior MS/MS results** because you want to combine the results of this search with the results of the previous one. You mark this check box only if you feel you have searched with incorrect settings or if for any other reason you want to start over.

Steps	Detailed Instructions	Comments
8 Start the search.	**a** Click the **Start Search** button. **b** View search progress in the area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• For the same data set and database, homology searches take longer than identity searches because there are more permutations to consider. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

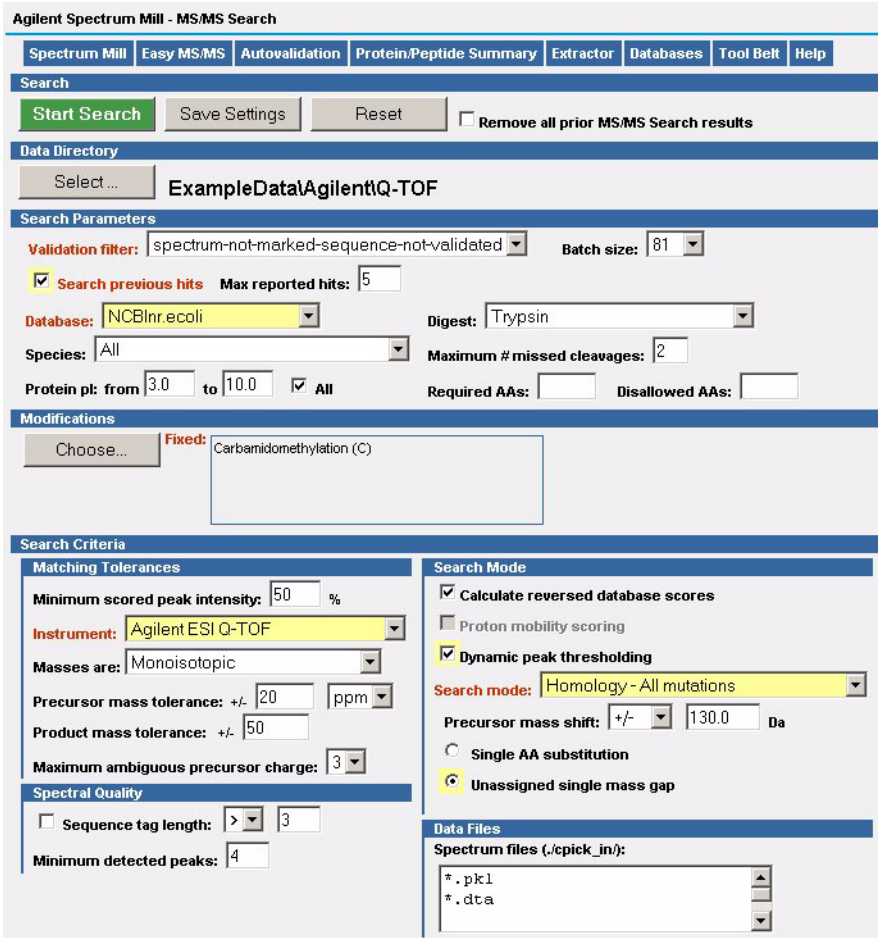


Figure 54 MS/MS Search settings for search of previous hits for unknown modifications

Exercise 3. Autovalidate the highest-quality results

After you have completed the database search, you validate the good results. Validation means that you accept that the matches are correct. The software adds the newly validated results to the ones you validated previously.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 55 on page 157.

2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed the data extraction and MS/MS Search for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Keep the defaults for the other settings, as shown in Figure 55 on page 157. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• The software adds these validated results to the ones you validated previously. • Unlike the processing you did in Chapter 4, here you do not apply the pI filter during autovalidation. To properly calculate the peptide pI, the pK of the modified amino acid must be specified in **smconfig.std.xml** or **smconfig.custom.xml**. Because you searched for unknown modifications, the software has no way to correlate the modification with a pK and so cannot calculate the correct pI of the modified peptides.

Steps	Detailed Instructions	Comments
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Keep the defaults for the other settings. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• The software adds these validated results to the ones you validated previously. • Again, you do not apply the peptide pI filter here.

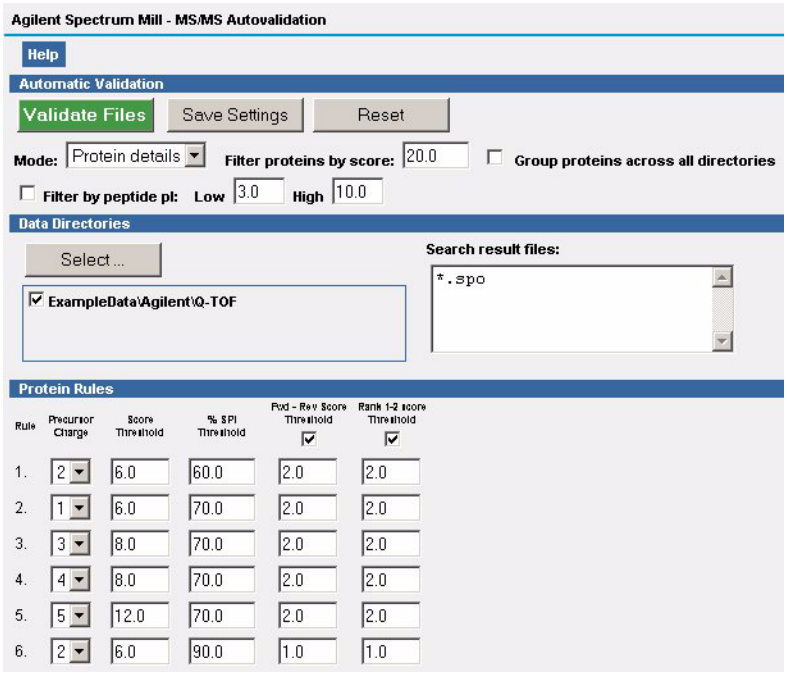


Figure 55 Autovalidation form, Protein details mode

Exercise 4. Examine the list of validated peptides

In this exercise, you examine the list of peptides you have validated from both the identity mode search and the mass gap search.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• From the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other Spectrum Mill pages.
2 Check that your **Data Directory** is set to **Q-TOF**.	• If you have just performed autovalidation for this data set, your **Q-TOF** data directory should already be set correctly. If not, click the **Select...** button to select the **Q-TOF** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Set the **Validation preset** to **none**.		

Steps	Detailed Instructions	Comments
6 Set the **Sort** parameters such that the unknown modifications are grouped at the top and/or bottom of the results table.	• For **Sort peptides by**, select **Delta mass shift**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	• By setting **Score** and **% SPI** to zero, you ensure that all valid data is selected, regardless of score.
8 Set the **Review Fields**.	**a** Mark the check box **Rank 1-2 score**. **b** To see the modifications, mark the check boxes for **Variable sites** and **Delta mass**. **c** For others, keep the default settings. **d** Check to see that your form looks like that in Figure 56 on page 160. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **Fwd-Rev score** and **Rank 1-2 score** help you rule out false positives. Small differences between forward and reversed scores, or between rank 1 and rank 2 scores, may indicate an incorrect identification. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		

Steps	Detailed Instructions	Comments
10 Examine the summary report.	**a** Check that your results are similar to those in Figure 57 on page 161 and Figure 58 on page 162. Since the public databases are constantly updated, your results may vary slightly. **b** Check that the data display includes the additional headings **Variable Sites**, **MH^+ Mass Shift (Da)**, and **MH^+ Error (ppm)**. **c** Note that the columns for **Variable Sites** and **MH^+ Mass Shift (Da)** summarize the modifications. To see results of the mass gap search, look for **Unknown** in the column for **Variable Sites**.	• This summary includes results validated from both identity mode and homology mode/mass gap searches. • Because you sorted by **Delta mass shift**, the software groups the results of the mass gap search at the top and bottom of the results table. • For a mass gap search, the column for **MH^+ Error (ppm)** shows the mass error relative to the unmodified peptide. For a variable modifications search (where the identity and mass of the modification are known), this column displays the mass error relative to the modified peptide.

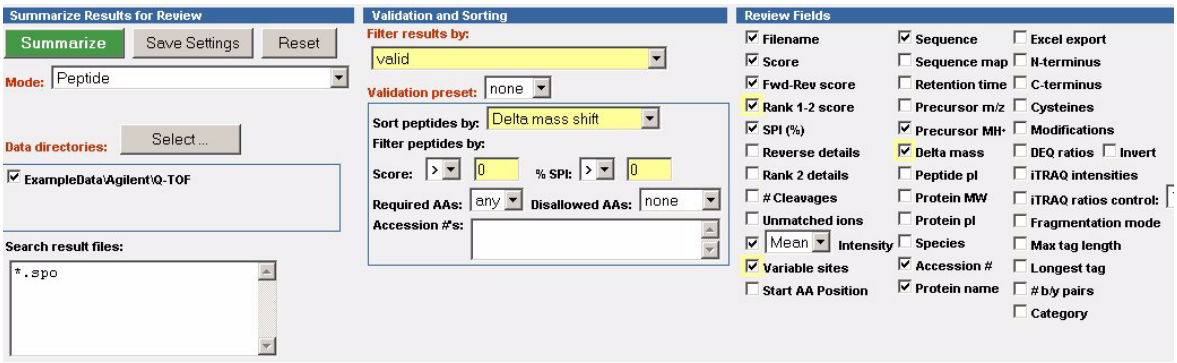
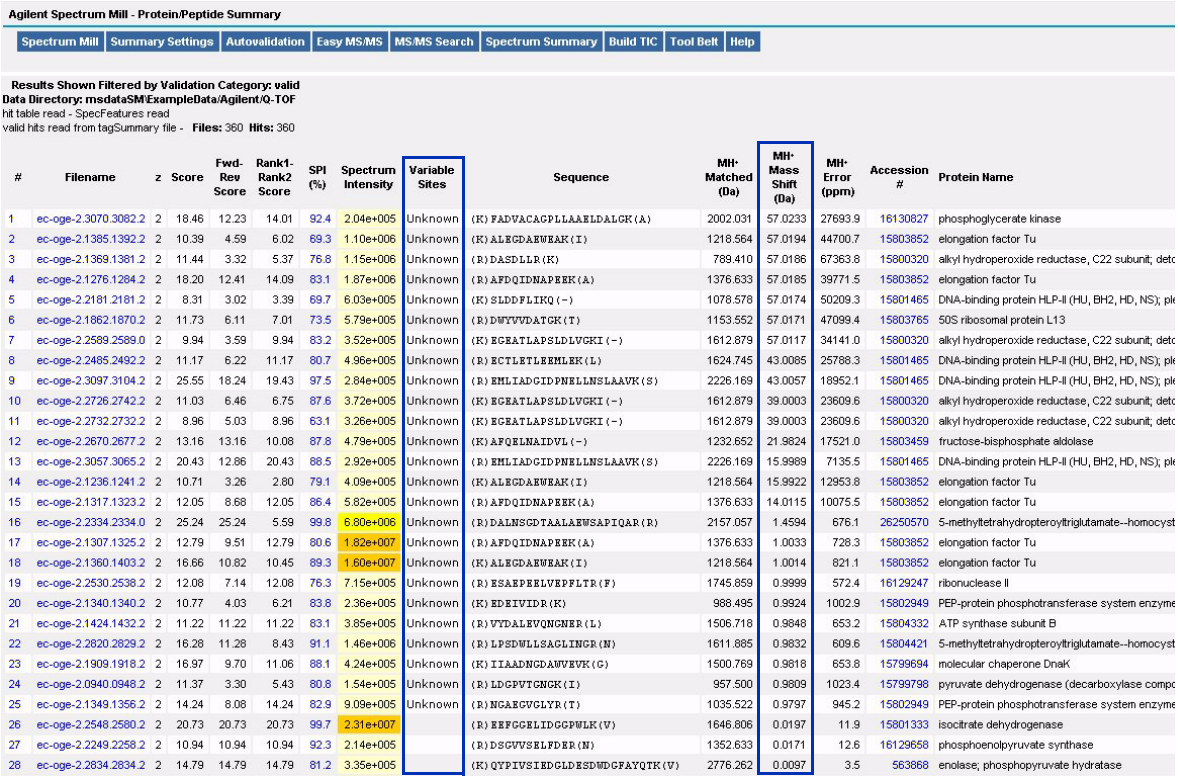


Figure 56 Settings to display results of mass gap search



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData/Agilent/Q-TOF
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 360 **Hits:** 360

#	Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	Spectrum Intensity	Variable Sites	Sequence	MH+ Matched (Da)	MH+ Mass Shift (Da)	MH+ Error (ppm)	Accession #	Protein Name
1	ec-oge-2.3070.3082.2	2	18.46	12.23	14.01	92.4	2.04e+005	Unknown	(K)FADVACAGPLLAAELDALGK(A)	2002.031	57.0233	27693.9	16130827	phosphoglycerate kinase
2	ec-oge-2.1385.1392.2	2	10.39	4.59	6.02	69.3	1.10e+006	Unknown	(K)ALEGDAEWEAK(I)	1218.564	57.0194	44700.7	15803852	elongation factor Tu
3	ec-oge-2.1369.1381.2	2	11.44	3.32	5.37	76.8	1.15e+006	Unknown	(R)DASDLLR(K)	789.410	57.0186	67363.8	15800320	alkyl hydroperoxide reductase, C22 subunit; detc
4	ec-oge-2.1276.1284.2	2	18.20	12.41	14.09	83.1	1.87e+006	Unknown	(R)AFDQIDNAPEEK(A)	1376.633	57.0185	39771.5	15803852	elongation factor Tu
5	ec-oge-2.2181.2181.2	2	8.31	3.02	3.39	69.7	6.03e+005	Unknown	(K)SLDDFLIKQ(-)	1078.578	57.0174	50209.3	15801465	DNA-binding protein HLP-II (HU, BH2, HD, NS); ple
6	ec-oge-2.1862.1870.2	2	11.73	6.11	7.01	73.5	5.79e+005	Unknown	(R)DWYVVDATGK(T)	1153.552	57.0171	47099.4	15803765	50S ribosomal protein L13
7	ec-oge-2.2589.2589.0	2	9.94	3.59	9.94	83.2	3.52e+005	Unknown	(K)EGEATLAPSLDLVGKI(-)	1612.879	57.0117	34141.0	15800320	alkyl hydroperoxide reductase, C22 subunit; detc
8	ec-oge-2.2485.2492.2	2	11.17	6.22	11.17	80.7	4.96e+005	Unknown	(R)ECTLETLEEMLEK(L)	1624.745	43.0085	25788.3	15801465	DNA-binding protein HLP-II (HU, BH2, HD, NS); ple
9	ec-oge-2.3097.3104.2	2	25.55	18.24	19.43	97.5	2.84e+005	Unknown	(R)EMLIADGIDPNELLNSLAAVK(S)	2226.169	43.0057	18952.1	15801465	DNA-binding protein HLP-II (HU, BH2, HD, NS); ple
10	ec-oge-2.2726.2742.2	2	11.03	6.46	6.75	87.6	3.72e+005	Unknown	(K)EGEATLAPSLDLVGKI(-)	1612.879	39.0003	23609.6	15800320	alkyl hydroperoxide reductase, C22 subunit; detc
11	ec-oge-2.2732.2732.2	2	8.96	5.03	8.96	63.1	3.26e+005	Unknown	(K)EGEATLAPSLDLVGKI(-)	1612.879	39.0003	23609.6	15800320	alkyl hydroperoxide reductase, C22 subunit; detc
12	ec-oge-2.2670.2677.2	2	13.16	13.16	10.08	87.8	4.79e+005	Unknown	(K)AFQELNAIDVL(-)	1232.652	21.9824	17521.0	15803459	fructose-bisphosphate aldolase
13	ec-oge-2.3057.3065.2	2	20.43	12.86	20.43	88.5	2.92e+005	Unknown	(R)EMLIADGIDPNELLNSLAAVK(S)	2226.169	15.9989	7135.5	15801465	DNA-binding protein HLP-II (HU, BH2, HD, NS); ple
14	ec-oge-2.1236.1241.2	2	10.71	3.26	2.80	79.1	4.09e+005	Unknown	(K)ALEGDAEWEAK(I)	1218.564	15.9922	12953.8	15803852	elongation factor Tu
15	ec-oge-2.1317.1323.2	2	12.05	8.68	12.05	86.4	5.82e+005	Unknown	(R)AFDQIDNAPEEK(A)	1376.633	14.0115	10075.5	15803852	elongation factor Tu
16	ec-oge-2.2334.2334.0	2	25.24	25.24	5.59	99.8	6.80e+006	Unknown	(R)DALNSGDTAALAEWSAPIQAR(R)	2157.057	1.4594	676.1	26250570	5-methyltetrahydropteroyltriglutamate--homocyst
17	ec-oge-2.1307.1325.2	2	12.79	9.51	12.79	80.6	1.82e+007	Unknown	(R)AFDQIDNAPEEK(A)	1376.633	1.0033	728.3	15803852	elongation factor Tu
18	ec-oge-2.1360.1403.2	2	16.66	10.82	10.45	89.3	1.60e+007	Unknown	(K)ALEGDAEWEAK(I)	1218.564	1.0014	821.1	15803852	elongation factor Tu
19	ec-oge-2.2530.2538.2	2	12.08	7.14	12.08	76.3	7.15e+005	Unknown	(R)ESAEPEELVEPFLTR(F)	1745.859	0.9999	572.4	16129247	ribonuclease II
20	ec-oge-2.1340.1340.2	2	10.77	4.03	6.21	83.8	2.36e+005	Unknown	(K)EDEIVIDR(K)	988.495	0.9924	1002.9	15802949	PEP-protein phosphotransferase system enzyme
21	ec-oge-2.1424.1432.2	2	11.22	11.22	11.22	83.1	3.85e+005	Unknown	(R)VYDALEVQNGNER(L)	1506.718	0.9848	653.2	15804332	ATP synthase subunit B
22	ec-oge-2.2820.2829.2	2	16.28	11.28	8.43	91.1	1.46e+006	Unknown	(R)LPSDWLLSAGLINGR(N)	1611.885	0.9832	609.6	15804421	5-methyltetrahydropteroyltriglutamate--homocyst
23	ec-oge-2.1909.1918.2	2	16.97	9.70	11.06	88.1	4.24e+005	Unknown	(K)IIAADNGDAWVEVK(G)	1500.769	0.9818	653.8	15799694	molecular chaperone DnaK
24	ec-oge-2.0940.0948.2	2	11.37	3.30	5.43	80.8	1.54e+005	Unknown	(R)LDGPVTGNGK(I)	957.500	0.9809	1023.4	15799798	pyruvate dehydrogenase (decarboxylase compo
25	ec-oge-2.1349.1356.2	2	14.24	8.08	14.24	82.9	9.09e+005	Unknown	(R)NGAEGVGLYR(T)	1035.522	0.9797	945.2	15802949	PEP-protein phosphotransferase system enzyme
26	ec-oge-2.2548.2580.2	2	20.73	20.73	20.73	99.7	2.31e+007		(R)EEFGGELIDGGPWLK(V)	1646.806	0.0197	11.9	15801333	isocitrate dehydrogenase
27	ec-oge-2.2249.2258.2	2	10.94	10.94	10.94	92.3	2.14e+005		(R)DSGVVSELFDER(N)	1352.633	0.0171	12.6	16129658	phosphoenolpyruvate synthase
28	ec-oge-2.2834.2834.2	2	14.79	14.79	14.79	81.2	3.35e+005		(K)QYPIVSIEDGLDESDWDGFAYQTK(V)	2776.262	0.0097	3.5	563868	enolase; phosphopyruvate hydratase

Figure 57 Mass gap search results, top

335	ec-oge-2.1568.1600.2	2	17.17	8.94	12.20	88.6	6.02e+006		(R)LTPDADDTVLVK(W)	1286.684	-0.0129	-10.0	15800310	2,3-dihydro-2,3-dihydroxybenzoate synthetase,
336	ec-oge-2.2505.2524.2	2	19.36	19.36	19.36	91.4	1.05e+006		(R)IEVPEIGEEVIEIK(A)	1596.873	-0.0150	-9.4	15803711	transcription elongation factor NusA
337	ec-oge-2.2780.2786.2	2	21.02	21.02	21.02	95.5	3.18e+005		(K)VALQDAGLSVSDIDDVILVGGQTR(M)	2441.288	-0.0157	-6.4	15799694	molecular chaperone DnaK
338	ec-oge-2.2833.2833.2	2	10.25	10.25	10.25	80.3	2.36e+005		(K)GLPADVVPGDILLLDDGR(V)	1834.991	-0.0200	-10.9	16129807	pyruvate kinase II
339	ec-oge-2.1300.1300.2	2	9.86	9.86	4.57	84.0	1.65e+006	Unknown	(K)QAFDDEELK(V)	1094.500	-0.0295	-27.0	15799684	threonine synthase
340	ec-oge-2.2853.2853.2	2	10.42	4.41	10.42	74.1	1.60e+006	Unknown	(K)ESVEAILGETLDLPK(E)	1613.863	-0.9947	-616.7	15799684	threonine synthase
341	ec-oge-2.2490.2497.2	2	11.24	6.26	11.24	78.0	2.84e+005	Unknown	(R)DVNVFCAPYDLVK(T)	1539.751	-3.9997	-2604.4	16131079	UDP-N-acetylglucosamine 1-carboxyvinyltransfe
342	ec-oge-2.1578.1578.2	2	11.26	5.40	5.46	74.0	3.15e+005	Unknown	(R)LTPDADDTVLVK(W)	1286.684	-4.0811	-3181.9	15800310	2,3-dihydro-2,3-dihydroxybenzoate synthetase,
343	ec-oge-2.1237.1243.2	2	11.06	4.64	11.06	73.6	3.27e+005	Unknown	(R)DLGLTDESK(W)	977.479	-4.9729	-5113.4	16129819	aspartyl-tRNA synthetase
344	ec-oge-2.0881.0881.2	2	13.75	8.93	10.71	78.5	2.98e+004	Unknown	(R)IPDGQAEDAR(K)	1071.507	-5.0378	-4723.8	15801728	aldehyde dehydrogenase, NAD-linked
345	ec-oge-2.3473.3484.2	2	11.09	3.64	6.34	78.0	1.85e+005	Unknown	(K)ILEAILLVWLDEAE(-)	1626.899	-8.0204	-4954.3	15803051	hypothetical protein Z3791
346	ec-oge-2.2197.2205.2	2	12.74	2.70	5.47	85.8	5.91e+005	Unknown	(R)CDILEPGTLQGYDRDPR(S)	2004.944	-17.0232	-8563.3	15804455	glutamine synthetase
347	ec-oge-2.2357.2357.2	2	11.93	2.70	11.93	99.1	7.17e+005	Unknown	(R)CTQELLFGK(G)	1095.550	-17.0268	-15787.1	16128895	aspartate aminotransferase, PLP-dependent
348	ec-oge-2.2182.2194.2	2	20.96	20.96	20.96	82.4	6.29e+005	Unknown	(R)QIVLNCGEEPSVVANTVK(G)	1957.006	-17.0270	-8776.9	15804735	chaperonin GroEL
349	ec-oge-2.3133.3162.2	2	18.84	10.07	15.02	92.3	4.53e+005	Unknown	(R)QSLGGLIEAYEAVAR(R)	1576.833	-17.0278	-10916.6	15802999	phosphoribosylaminoimidazole-succinocarboxam
350	ec-oge-2.1825.1832.2	2	13.25	4.81	4.65	78.7	1.02e+006	Unknown	(K)QALELPR(E)	826.478	-17.0292	-21038.0	15800166	trigger factor
351	ec-oge-2.1177.1183.2	2	16.02	6.96	2.83	87.8	9.68e+005	Unknown	(R)QAIVEQER(L)	972.511	-17.0296	-17823.0	15804421	5-methyltetrahydropteroyltriglutamate--homocyst
352	ec-oge-2.2341.2350.2	2	17.97	13.51	14.59	81.8	2.89e+005	Unknown	(R)CVEKDIDVAAESLK(V)	1576.789	-17.0300	-10918.3	16128068	2-isopropylmalate synthase
353	ec-oge-2.1339.1347.2	2	21.51	21.51	21.51	86.3	6.61e+005	Unknown	(R)QQIEEATSDYDREK(L)	1711.777	-17.0302	-10048.8	15804735	chaperonin GroEL
354	ec-oge-2.0977.0980.2	2	16.86	3.63	4.71	93.1	1.51e+005	Unknown	(K)QAQQELR(D)	872.458	-17.0305	-19908.8	15800768	phosphoserine aminotransferase
355	ec-oge-2.2174.2174.2	2	20.67	14.63	16.62	84.8	3.63e+005	Unknown	(R)QLDPLVVGQEHYDTAR(G)	1840.919	-17.0305	-9337.5	15804332	ATP synthase subunit B
356	ec-oge-2.1256.1263.2	2	14.04	14.04	10.20	79.2	1.11e+006	Unknown	(K)QQYGEGLAR(S)	1021.506	-17.0308	-16954.9	16131330	leucine transporter subunit
357	ec-oge-2.0847.0854.2	2	13.98	6.52	6.74	86.5	1.10e+005	Unknown	(R)QAEKEGQEPDPR(V)	1383.650	-17.0316	-12462.6	26250729	phosphoenolpyruvate carboxylase
358	ec-oge-2.2310.2310.2	2	15.26	15.26	15.26	86.3	2.87e+005	Unknown	(R)QQFAQVTNPPIDPLR(E)	1723.913	-17.0597	-9994.8	16131102	glutamate synthase, large subunit
359	ec-oge-2.1526.1526.2	2	7.69	7.69	7.69	70.3	1.17e+005	Unknown	(R)AFDQIDNAPEEK(A)	1376.633	-18.0155	-13260.2	15803852	elongation factor Tu
360	ec-oge-2.1902.1905.2	2	16.99	13.14	16.99	81.4	1.88e+005	Unknown	(-)MSEQHAQGADAVVDLNNELK(T)	2169.024	-89.0289	-42802.5	16130792	lysine tRNA synthetase, constitutive

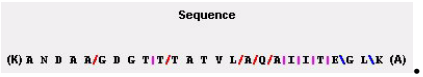
360 Files listed.

Figure 58 Mass gap search results, bottom

Exercise 5. Examine details of search results to determine modification sites

In this exercise, you examine details that help you localize the site of the peptide modification.

Steps	Detailed Instructions	Comments
1 If you have not already done so, display a Protein/Peptide Summary report.	• Follow steps in "Exercise 4. Examine the list of validated peptides" on page 158.	
2 Display MS/MS Search results.	**a** Under the **Filename** header, click the first entry. **b** Check the bottom of the browser window, where you see the annotated spectrum, as shown in Figure 59 on page 164. **c** Note the presence of star-ions (**b^*_2**, **b^*_3**, **b^*_4**, **b^*_5**). These are the complements of the detected y-ions, and they contain the modification. In other words, they are the b-ions plus the mass shift.	• Note that **MH^+ Mass Shift** shows an addition of 57 Da.
3 Expand the section of the page that shows the MS/MS Search results.	**a** To adjust the size of this section, rest your mouse over the split line. **b** When the pointer becomes a double-sided arrow, drag the arrow to move the split line.	
4 Scroll up to see more results.	**a** Scroll up until you see the heading for **Detailed Results**. **b** Check that your display resembles that in Figure 60 on page 165.	
5 Examine the table where the ions are assigned.	**a** Note that the table shows y-ions and b*-ions, but no b-ions. This means that the modification is on the N-terminus.	• The addition of 57 Da is likely carbamidomethylation of the N-terminus, an artifact of cysteine alkylation.

Steps	Detailed Instructions	Comments
6 Examine the sequence of the database hit.	**a** Look above the table of ion assignments to see a sequence or list of sequences. **b** Note the color-coded lines in the sequence below (from an identity-mode search) and read the comment to the right in this table.  **c** Examine the sequence in Figure 60 on page 165 to see the locations of identified y-ions in our example. (Again, there are no b-ions because the N-terminus is modified.)	• The color-coding of the amino acid sequence of the matched peptide from the database search shows: ◦ Red forward-slashes for locations of y-ions ◦ Blue backslashes for locations of b-ions (not b*-ions) ◦ Magenta pipes (vertical lines) for locations of both b- and y-ions • For mass gap results, the color-coded sequence helps you to identify the site of a modification by highlighting the existence of surrounding b/y ions.
7 (Optional) Learn more details about this view of the MS/MS Search results.	**a** See "Exercise 13. (Optional) Examine details of search results" on page 55. **b** See the information about manual review and validation in Chapter 1 of the *Application Guide*.	•

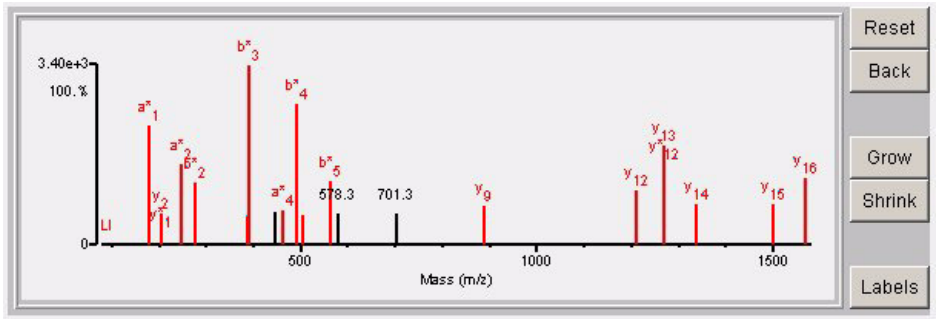


Figure 59 Spectrum from MS/MS Search results, showing star-ions

Detailed Results

Rank	Score	SPI (%)	BCS	# Unmatched Ions	Variable sites	Sequence	MH+ Calculated (Da)	MH+ Error (Da)	MH+ Error (ppm)	Protein MW/pI (Da)	Species	Accession #	MS-Digest Index #	Protein Name
1	18.46	92.4	9	4/20	Unknown	(K) F A D V/R/C/R/G/P L L/R A E L/D/R L/G K (A)	2002.0314	57.0233	27693.9	41118.4/5.08	Escherichia coli K12	16130827	6201	phosphoglycerate kinase
1	18.46	92.4	9	4/20	Unknown	(K) F A D V/R/C/R/G/P L L/R A E L/D/R L/G K (A)	2002.0314	57.0233	27693.9	41130.5/5.08	Escherichia coli O157:H7 EDL933	15803460	6676	phosphoglycerate kinase
1	18.46	92.4	9	4/20	Unknown	(K) F A D V/R/C/R/G/P L L/R A E L/D/R L/G K (A)	2002.0314	57.0233	27693.9	43298.0/5.23	Escherichia coli CFT073	26249339	11745	phosphoglycerate kinase

Fragment-ion (m/z)	86.100	177.099	204.127	248.134	276.133	388.254	391.154	444.201	462.239	490.225	503.292	561.255	578.301	701.336	887.486	1210.679	1267.722	1338.751	149
Frac. Inten.(% of TIC)	0.11	9.85	2.58	6.72	5.10	2.32	14.90	2.69	2.86	11.71	2.46	5.31	2.65	2.58	3.24	4.46	8.22	3.35	
Rel. Inten.(% of BP)	0.74	66.10	17.29	45.13	34.22	15.57	100.00	18.09	19.19	78.58	16.51	35.62	17.77	17.31	21.76	29.95	55.17	22.46	
Score	-0.01	0.50	1.50	0.50	1.00	1.50	1.00	-0.18	0.50	1.00	1.50	1.00	-0.18	-0.17	1.50	1.50	1.50	1.50	
Ion-type	LI	a^*_1	y_2	a^*_2	b^*_2	y_4	b^*_3		a^*_4	b^*_4	y_5	b^*_5			y_9	y_{12}	y_{13}	y_{14}	
Delta ppm	35.7	-1999946.6	-37.6	-1999954.9	-1999978.8	-4.0	-1999968.3		-2000002.3	-1999982.9	19.9	-1999970.4			2.7	-21.0	-2.7	-9.1	
			y^*_1 -1999935.2														y^*_{12} -1999995.6		

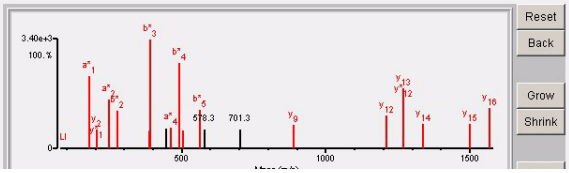


Figure 60 MS/MS Search results, showing the spectrum, table of fragment-ion assignments, and sequence

6
Processing ICAT Data

Exercise 1. (Optional) Transfer ICAT data file to Spectrum Mill server 169
Exercise 2. (Optional) Extract spectra from ICAT data 171
Exercise 3. Create indices for a user database 176
Exercise 4. Run database search on ICAT data 178
Exercise 5. Autovalidate the ICAT data 183
Exercise 6. Summarize valid results in peptide mode 186
Exercise 7. Summarize valid results in protein mode 190

This chapter provides familiarization exercises to process an example ICAT MS/MS data file with the Agilent Spectrum Mill MS Proteomics Workbench. This chapter elaborates on the ICAT chapter of the companion *Spectrum Mill MS Proteomics Workbench Application Guide*. These instructions are intended as an introduction to the basic operation of the software; you are encouraged to customize these steps for your samples and studies.

In general, you process ICAT data much the same as any other ion trap MS/MS data. To ensure accurate processing and to facilitate ICAT quantitation, there are some specific parameters you set in Data Extractor, MS/MS Search, Protein/Peptide Summary, and Spectrum Summary. Note that ICAT calculations are not supported for data that uses the generic Data Extractor.

While this chapter focuses specifically on quantitation with the original ICAT D_0/D_8 reagent, the same principles apply for other isotopically labeled modifications that are used for relative quantitation. You process the following modifications almost identically to ICAT:

- cICAT -^{12}C, cICAT-^{13}C, and cICAT-mix
- C-terminal methyl ester-D_0, methyl ester-D_3, and methyl ester-mix
- N-terminal propionyl-D_0, propionyl-D_5, and propionyl-mix
- C-terminal $^{16}O/^{18}O$

You can also process other types of labels with the Spectrum Mill workbench:

- iTRAQ reagents (simultaneous quantitation of up to four different cell states based on low-mass MS/MS signature ions)
- Metabolic labels:
  - SILAC 2 (Arg 0-6Da, Lys 0-8Da)-mix
  - SILAC 3 (Arg 0-6-10Da)-mix
  - ^{15}N and $^{14}N/^{15}N$ mix

The iTRAQ tags are different from ICAT because they use low-mass MS/MS signature ions rather than isotopes as labels. The software processes the data differently; for details, see the *Application Guide*.

The metabolic labels are also a little different than ICAT. With most modifications, you can specify only one fixed or variable modification per amino acid. With metabolic modifications, you can specify a fixed or variable modification to the same amino acid that is modified by the metabolic modification. For example, if a SILAC label modifies K, you can still select guanidination of K as a fixed modification and carbamylation of K as a variable modification. The software applies the proper isotopes for each type of modification.

There is one additional difference that applies only to the $^{14}N/^{15}N$ mix. With the nitrogen isotope labels, quantitation begins at the Protein/Peptide Summary level rather than at the Data Extractor level. Unlike the other isotopic labels, both the ^{14}N and the ^{15}N peptides must have been subjected to MS/MS, and the MS/MS Search results must indicate the same sequence and charge.

There are two types of labels that are used for quantitation and for which data can originate with the generic Data Extractor:

- iTRAQ
- $^{14}N/^{15}N$ mix.

Exercise 1. (Optional) Transfer ICAT data file to Spectrum Mill server

In this exercise, you move the ICAT data into the appropriate directory on the Spectrum Mill server.

NOTE The ICAT example files may have been installed on the server at the time the Spectrum Mill software was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

Steps	Detailed Instructions	Comments
1 Copy the example data directory, **X:\ExampleData\msdataSM\ExampleData\ICAT**, from the *Spectrum Mill Example Data CD* onto your client PC.	• Copy the **ICAT** directory from the *Spectrum Mill Example Data CD* to any directory on your client PC.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • This exercise simulates the likely laboratory scenario of file transfer from a client or instrument PC to the server; you may transfer files directly from the CD to the server if you prefer.
2 Create the **msdataSM\ExampleData\ICAT** folder on the Spectrum Mill server.	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In **msdataSM**, create **ExampleData\ICAT**. You need only one folder because this exercise uses a single sample that contains a mix of ICAT labels.	• If you don't know how to find your Spectrum Mill file system, ask the person who installed your software. • **Do not include spaces or parentheses in your directory name.** • When you process your own samples, remember to set up a separate folder for each sample. Each folder should contain all sample fractions. • You may create up to ten folders between **msdataSM** and your data files, but shorter path lengths reduce memory usage, especially for large data sets.

Steps	Detailed Instructions	Comments
3 Copy the example files to the new directory.	**a** Copy or move the contents of the **ICAT** directory from your client PC to the new folder on the server PC. **b** Be sure to include both the **ICAT_Mix.RAW** file and the **cpick_in** subdirectory.	• For convenience, these data are provided pre-extracted. The **cpick_in** subdirectory contains the extracted spectral files. • When you process your own data, remove any spaces or parentheses in the data file names. For example, change the file **my sample.d** to **my_sample.d**.
4 Make sure you have both read and write permissions for the data folder you just created.	**a** Right-click the **ICAT** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked. **c** Click the **Security** tab. **d** Make sure all user groups have full permissions. **e** Click **OK**.	If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors.

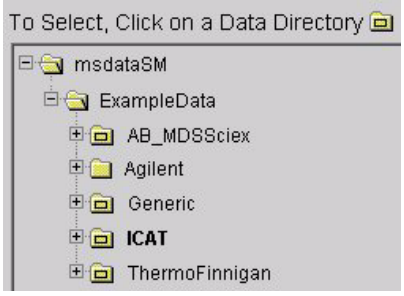
Exercise 2. (Optional) Extract spectra from ICAT data

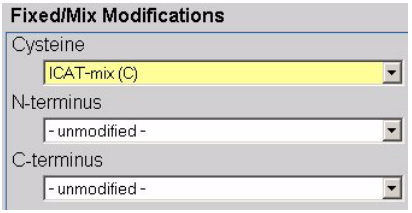
The Spectrum Mill Data Extractor preprocesses raw data files to extract high-quality spectra for database searches. When you use ICAT reagents, you select the appropriate cysteine modifications in the Data Extractor program. For the original ICAT reagent, these include **ICAT-D0**, **ICAT-D8**, and **ICAT-mix**. For the newer cleavable ICAT reagent, these include **cICAT-C12**, **cICAT-C13**, and **cICAT-mix**. If your sample contains a mixture of light and heavy labels and you want the ratio calculated, you select **ICAT-mix** or **cICAT-mix**.

When the cysteine modification is set to a mixture of ICAT labels, Data Extractor calculates peak areas for a number of additional parallel EICs, one of which represents the other member of the ICAT pair. The reason Data Extractor calculates a *number* of additional EICs is because at this point it does not know whether it has extracted a cysteine-containing residue, a light member of the pair, or a heavy member of the pair. It also does not know how many cysteines are present, and it does not know the charge state of the precursor ions. All the parallel EIC areas are stored until the Spectrum Mill workbench interpretation is complete. Then the correct ones are retrieved for ICAT calculations.

To see the intensities of these parallel EICs, go to the Spectrum Summary page and mark the **MS L/H EIC Intensity** check box under **Review Fields**. Since you typically use the Spectrum Summary page to display spectra that have not been interpreted, you do not see the light/heavy ratios calculated in this table, nor should you attempt to calculate them from these data. Instead, use the calculations from the Protein/Peptide Summary page, as described later in this chapter.

NOTE Since the ICAT data are provided pre-extracted, this exercise is optional. You need the optional Extractor License for Thermo Ion Trap Data to do this exercise.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link. 	
2 Select the **msdataSM\ExampleData\ICAT** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **ICAT** to select that data file directory.  **c** Make sure that the name of the **ICAT** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 61 on page 175.	• Directories are identified by different types of icons: ◦ Plain folders indicate directories that do not have data files directly beneath them. ◦ Folders with rectangles indicate data directories. ◦ Folders with line spectra (bar graphs) indicate data files. • Try clicking the names of each type to see which turn **bold**, indicating that they are selectable. • The software remembers your data file selection when you go to other Spectrum Mill forms. • If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **ICAT-mix**. The sample contains a mix of D_0 and D_8 ICAT labels.  **c** Click **OK.** The name of the modification appears as in Figure 61 on page 175.	• For other samples, you may select **ICAT-D0, ICAT-D8**, **cICAT-C12**, **cICAT-C13**, **cICAT-mix**, etc. • This step is essential if you want to have the software calculate the ICAT ratio.
4 Set other parameters as shown in Figure 61 on page 175.	**a** Keep the defaults whenever appropriate. This example uses all default settings, except for **Data Directory** and **Modifications**. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	• You use the **Sequence tag length** to filter out noisy spectra. This is the longest sequence of amino acids that is represented by the fragments in a spectrum. • If you set this to **>1**, you ensure that all possible good spectra are extracted. You have opportunities to set more stringent requirements later when you perform the database search. • When you process your own samples, if you wish to perform quantitation for labels with small mass differences (~4 Da) between the light and heavy versions, change **Merge scans with same precursor m/z** to **+/-1.0 m/z** or lower.

Steps	Detailed Instructions	Comments
5 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** area to the right of the Data Extractor form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.
6 (Optional) View the **cpick_in** subdirectory to see the files Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\ICAT\cpick_in**. **b** Notice the new files created there. Each one represents an extracted spectrum from your raw data file.	File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous)

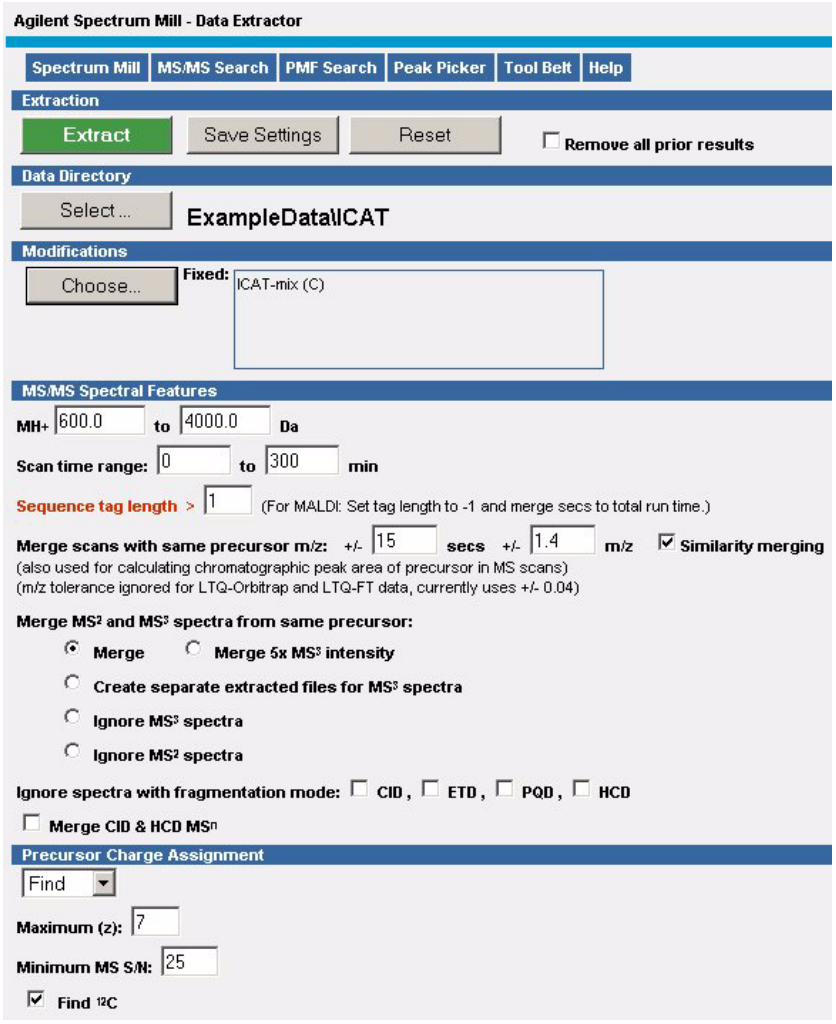
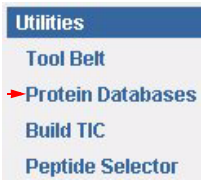

Agilent Spectrum Mill - Data Extractor
Spectrum Mill
MS/MS Search
PMF Search
Peak Picker
Tool Belt
Help
Extraction
Extract
Save Settings
Reset
Remove all prior results
Data Directory
Select ...
ExampleData\ICAT
Modifications
Choose...
Fixed:
ICAT-mix (C)
MS/MS Spectral Features
MH+ 600.0 to 4000.0 Da
Scan time range: 0 to 300 min
Sequence tag length > 1 (For MALDI: Set tag length to -1 and merge secs to total run time.)
Merge scans with same precursor m/z: +/- 15 secs +/- 1.4 m/z Similarity merging
(also used for calculating chromatographic peak area of precursor in MS scans)
(m/z tolerance ignored for LTQ-Orbitrap and LTQ-FT data, currently uses +/- 0.04)
Merge MS² and MS³ spectra from same precursor:
Merge
Merge 5x MS³ intensity
Create separate extracted files for MS³ spectra
Ignore MS³ spectra
Ignore MS² spectra
Ignore spectra with fragmentation mode: CID , ETD , PQD , HCD
Merge CID & HCD MSⁿ
Precursor Charge Assignment
Find
Maximum (z): 7
Minimum MS S/N: 25
Find ¹²C

Figure 61 Settings for Data Extractor with ICAT data

Exercise 3. Create indices for a user database

The Spectrum Mill workbench allows you to create and search user databases. In this exercise, you create indices for a user database so you can later search it.

Steps	Detailed Instructions	Comments
1 If necessary, copy the user database **NCBInr.stdmix** to your Spectrum Mill server.	**a** If the **NCBInr.stdmix** database is already installed in your Spectrum Mill database directory, skip to step 2. **b** If not, find the file **X:\Example Databases\NCBInr.stdmix** on your *Spectrum Mill Installation CD*. **c** Copy it to your database directory on your Spectrum Mill server.	• The database directory on the server likely has a name like **X:\seqDB**. • If you are not sure, ask your system administrator.
2 Make sure security permissions are set properly for the folder that contains the file **dbname.js**.	**a** Right-click the **SpectrumMill\millhtml\SM_js** folder and select **Properties**. **b** Click the **Security** tab. **c** Make sure all user groups have full permissions. **d** Click **OK**.	If the Spectrum Mill workbench cannot write to this folder, you may encounter errors.
3 Navigate to the Protein Databases page.	• From the Spectrum Mill home page, click the Protein Databases link. 	
4 Fill in the form as shown in Figure 62 on page 177.	**a** On the Protein Databases page, click the **Create indices for new database** option. (This is the default.) **b** Set **Newly downloaded database** to **NCBInr.stdmix.**	For **Newly downloaded database**, you type the exact database file name.

Steps	Detailed Instructions	Comments
5 Click the **Create Indices** button.	• After a short wait, check the bottom of the page for text that indicates that the software is creating the database indices.	These indices bear no resemblance to those used by another popular database search program.

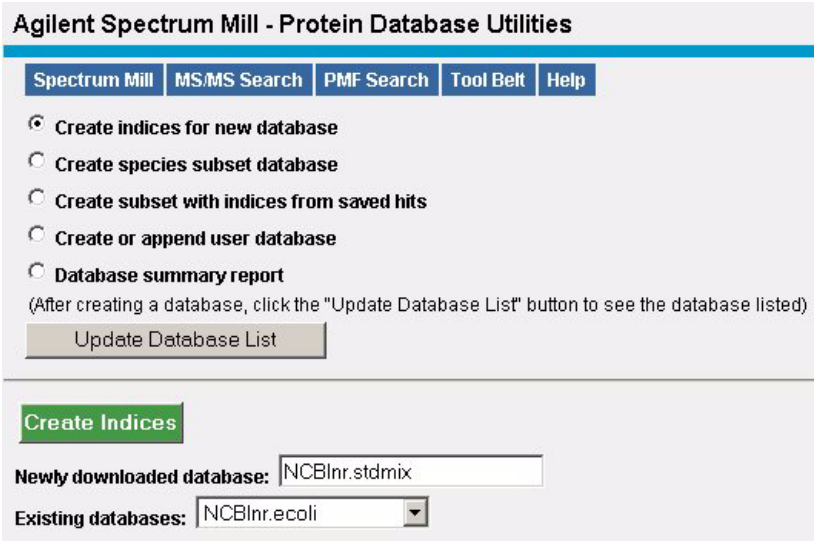


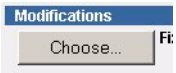
Figure 62 Create indices for user database

Exercise 4. Run database search on ICAT data

After you have extracted your spectra and created indices for your user database, you are ready to search each spectrum against a protein or DNA database. As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal to identify as many spectra as possible. These exercises illustrate only a single identity mode search. In identity mode, the spectra must be consistent with the database sequence. When you process your own data, you might also search in variable modifications or homology mode, which allow for modifications relative to the database.

When you use ICAT reagents, you select ICAT-related cysteine modifications in the MS/MS Search program. If you select **ICAT-mix**, each spectrum is searched twice – once for modification by the light reagent and once for modification by the heavy reagent. The results are merged as a single output.

Steps	Detailed Instructions	Comments
1 Navigate to the MS/MS Search page.	• Do one of the following: ◦ From the Protein Databases page, click the **MS/MS Search** button. ◦ From the Spectrum Mill home page, click the link to MS/MS Search. 	Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Protein Databases page to the MS/MS Search page.
2 Check that your **Data Directory** is set to **ICAT**.	• If you have just performed the data extraction, your **ICAT** data directory should already be set correctly. If not, click the **Select...** button to select the **ICAT** folder.	

Steps	Detailed Instructions	Comments
3 Select the "standard mix" database.	• For **Database** (under **Search Parameters**), select the **NCBInr.stdmix** database for which you just created indices.	If the database name does not appear, reload or refresh the MS/MS search form.
4 Choose the appropriate cysteine modification.	**a** Click the **Choose...** button near the middle of the form.  **b** Under the **Cysteine** heading, select **ICAT-mix**. **c** Click **OK.** The name of the modification appears as in Figure 63 on page 182.	• For other samples, you may select **ICAT-D0**, **ICAT-D8**, **cICAT-C12**, **cICAT-C13**, **cICAT-mix**, etc. • This step is essential to have each spectrum searched twice — once for modification by the light reagent and once for modification by the heavy reagent. The results are merged as a single output.
5 Choose the appropriate instrument.	• For **Instrument** (under **Matching Tolerances**), select **ESI ion trap**.	

Steps	Detailed Instructions	Comments
6 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **b** Mark the check box for **Proton mobility scoring**. **c** Mark the check box for **Dynamic peak thresholding**. **d** Verify that the **Search mode** is set to **Identity**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Protein mobility scoring** is a scoring enhancement for ion trap data that applies a bonus or penalty based on the expected relative intensities for MS/MS fragments from a given peptide sequence. When you process your own ion trap data, it is best to use this scoring, except with modifications that significantly alter the fragmentation patterns of peptides. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.
7 Set other parameters as shown in Figure 63 on page 182.	**a** Keep the defaults whenever appropriate. **b** Make sure **Search mode** is set to **Identity** for this round of processing. **c** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **d** Click a blue section divider bar to display help for that section of the form.	

Steps	Detailed Instructions	Comments
8 Start the search.	**a** Click the **Start Search** button **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • Search time varies depending on the size of the database searched. This search goes fairly quickly because you search a small user database. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

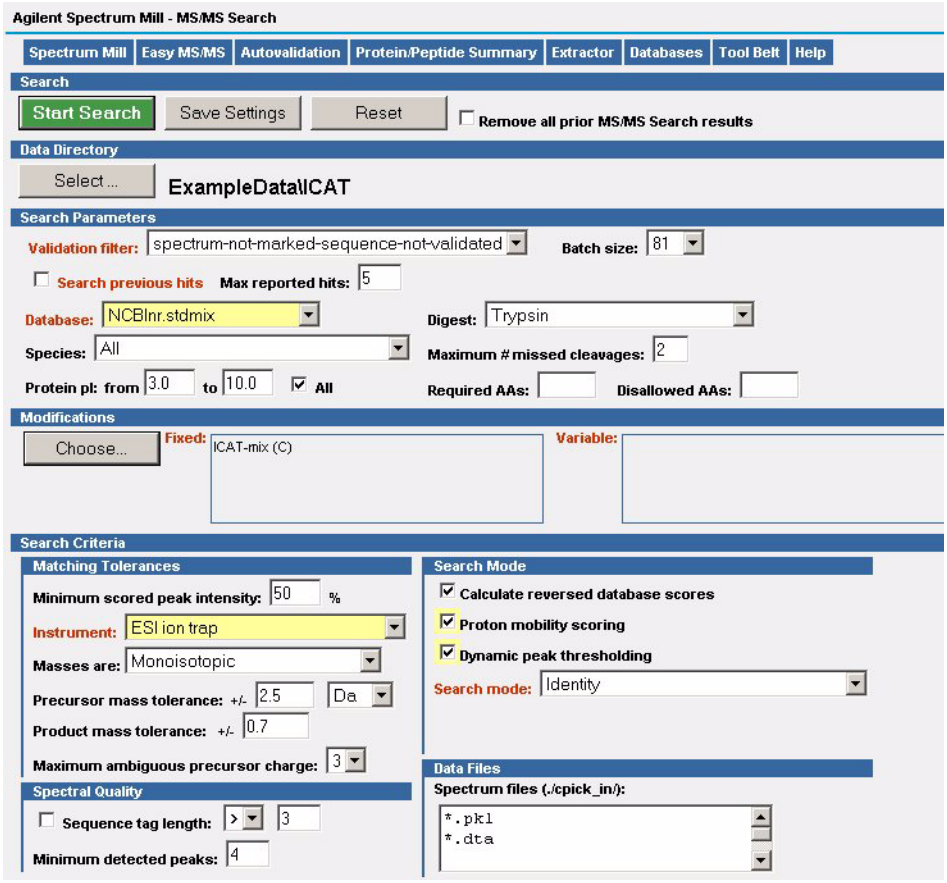


Figure 63 MS/MS Search settings for ICAT data

Exercise 5. Autovalidate the ICAT data

After you have completed a database search, you validate the good results. Validation means that you accept that the matches are correct.

The Spectrum Mill workbench provides a means to segregate search results that contain a valid interpretation of an MS/MS spectrum from those that do not. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

The Spectrum Mill workbench provides two ways to validate results. One way uses the Autovalidation page, and is totally automated. You use this method only to validate the highest-scoring results — those that do not require manual review. The other method of validating uses the Protein/Peptide Summary page for manual review and validation. This exercise describes autovalidation. To learn more about manual validation, see page 37 through page 44.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 64 on page 185.

2 Check that your **Data Directory** is set to **ICAT**.	• If you have just performed data extraction and MS/MS search with this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **ICAT** folder.	
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Keep the default scoring presets. **c** Click the **Validate Files** button. **d** Watch for a Spectrum Mill Validation Summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein. • Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20.

Steps	Detailed Instructions	Comments
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Keep the default scoring presets. **c** Click the **Validate Files** button. **d** Watch for a Spectrum Mill Validation Summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by peptide. Even if it finds only a single peptide for a protein, it validates the corresponding search results provided that the peptide score is high enough. • Using the default scoring, individual peptides must have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein.
5 Close the Autovalidation page.		

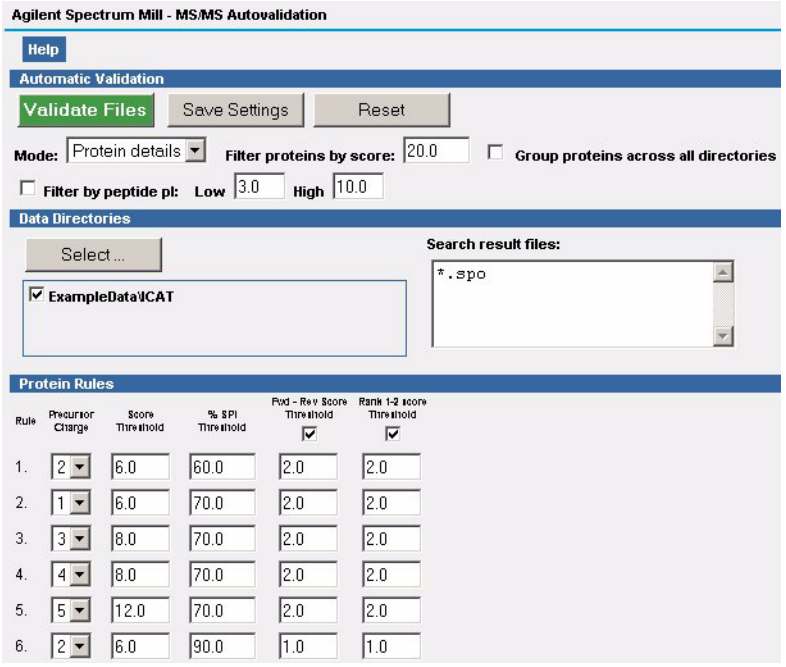


Figure 64 Autovalidation form with ICAT data

Exercise 6. Summarize valid results in peptide mode

In this exercise, you summarize the results that you just automatically validated. In this exercise, you summarize in peptide mode, while in the next exercise, you summarize in one of the protein modes.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page, if you are not already there.	• From the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other Spectrum Mill pages.
2 Check that your **Data Directory** is set to **ICAT**.	• If you have just performed autovalidation for this data set, your data directory should already be set correctly. If not, click the **Select...** button to select the **ICAT** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set the **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Leave the **Validation preset** set to the default of **none**.		• You use this setting when you want to summarize results rather than manually review and validate results. • If you want to manually review and validate these results, see page 37 through page 44.

Steps	Detailed Instructions	Comments
6 Set **Sort peptides by** to **Sequence**.		• The **Sort** parameters determine how the data are sorted in the results display. • In this case, you sort by sequence so that ICAT pairs are displayed together. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option appear. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	By setting to zero, you ensure that all valid data is selected, regardless of score.
8 Filter results so that only cysteine-containing peptides are shown.	• In the **Required AAs** box, select **C**.	The software can perform amino acid filtering at either the MS/MS Search level or the Protein/Peptide Summary level.
9 Set the **Review Fields**.	**a** Mark the **DEQ ratios** check box in the last column. **a** Mark the **Cysteines** check box as well. **b** Keep the other default settings. (Make sure the **Intensity** check box is marked.) **c** Check to see that your form looks like that in Figure 65 on page 188. Note that changes from defaults are highlighted in yellow.	• These settings specify the data fields that are displayed. • **DEQ ratios** are differential expression quantitation ratios — the ratios of the light/heavy pairs. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
10 Click the **Summarize** button.		

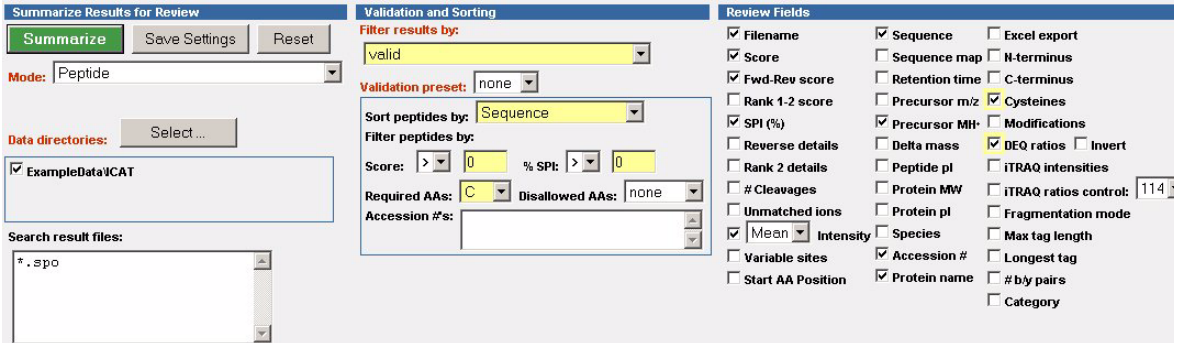


Figure 65 Settings to summarize valid ICAT results - peptide mode

Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\ICAT
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 38 **Hits:** 37

#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	Sequence	Cysteines	L/H	MH+ Matched (Da)	Accession #	Protein Name
1	ICAT_Mix.1064.1064.0	2	14.10	14.10	79.5	8.13e+006	(K)AGLCQTFVYGGCR(A)	ICAT-D0	0.891	2259.079	115114	Aprotinin
2	ICAT_Mix.0837.0842.2	2	13.92	13.92	71.2	1.03e+008	(K)CDEWSVNSVGK(I)	ICAT-D8	0.717	1673.811	4557871	transferrin
3	ICAT_Mix.0843.0847.2	2	15.44	15.44	73.4	8.62e+007	(K)CDEWSVNSVGK(I)	ICAT-D0	0.812	1665.761	4557871	transferrin
4	ICAT_Mix.1083.1083.2	2	9.16	9.16	81.7	2.94e+007	(K)CSTSSLLEACTFR(R)	ICAT-D8	0.266	2318.195	4557871	transferrin
5	ICAT_Mix.1095.1098.2	2	10.75	10.75	74.2	1.47e+007	(K)CSTSSLLEACTFR(R)	ICAT-D0	1.090	2302.095	4557871	transferrin
6	ICAT_Mix.1067.1067.0	3	11.61	11.61	97.3	5.54e+007	(K)DAIPENLPPLTADFAEDKDVCK(N)	ICAT-D8	n/c	2851.434	1351907	bovine serum albumin
7	ICAT_Mix.1068.1073.0	3	19.26	19.26	95.2	5.00e+007	(K)DAIPENLPPLTADFAEDKDVCK(N)	ICAT-D0	n/c	2843.384	1351907	bovine serum albumin
8	ICAT_Mix.1072.1072.0	3	11.36	11.36	96.8	6.29e+007	(K)DAIPENLPPLTADFAEDKDVCK(N)	ICAT-D8	n/c	2851.434	1351907	bovine serum albumin
9	ICAT_Mix.0863.0863.0	2	7.95	7.11	88.9	1.72e+007	(K)DDPHACYSTVFDK(L)	ICAT-D0	1.559	1939.857	1351907	bovine serum albumin
10	ICAT_Mix.0997.0997.0	2	12.58	12.58	77.4	5.52e+007	(K)DYELLCLDGTR(K)	ICAT-D8	0.492	1747.884	4557871	transferrin
11	ICAT_Mix.1002.1008.0	2	21.43	21.43	98.2	3.80e+007	(K)DYELLCLDGTR(K)	ICAT-D0	0.591	1739.834	4557871	transferrin
12	ICAT_Mix.1003.1003.2	2	12.83	12.83	74.9	5.84e+007	(K)DYELLCLDGTR(K)	ICAT-D8	0.497	1747.884	4557871	transferrin
13	ICAT_Mix.0844.0844.0	3	11.30	11.30	73.4	1.77e+007	(R)EGTCPEAPTDECKPVK(W)	ICAT-D8	n/c	2604.312	4557871	transferrin
14	ICAT_Mix.0922.0922.0	2	8.40	8.40	65.0	2.14e+007	(K)EYEATLEECCAK(D)	ICAT-D8	0.339	2289.121	1351907	bovine serum albumin
15	ICAT_Mix.0967.0967.0	2	11.35	11.35	83.7	4.36e+007	(R)FDEFFSEGCAPGSK(K)	ICAT-D8	0.416	1970.911	4557871	transferrin
16	ICAT_Mix.0972.0972.0	2	9.79	9.79	75.5	5.25e+007	(R)FDEFFSEGCAPGSK(K)	ICAT-D8	0.582	1970.911	4557871	transferrin
17	ICAT_Mix.1044.1044.0	3	10.84	10.84	85.2	1.64e+007	(K)LCMGSGLNLCEPNNK(E)	ICAT-D8	n/c	2493.273	4557871	transferrin
18	ICAT_Mix.1202.1202.2	2	13.19	13.19	93.6	5.31e+006	(R)MPCTEDYLSLILNR(L)	ICAT-D8	0.739	2118.088	1351907	bovine serum albumin
19	ICAT_Mix.1283.1288.0	3	18.59	18.59	88.3	9.24e+006	(R)SAGWNIPIGLLYCDLPEPR(K)	ICAT-D8	n/c	2564.349	4557871	transferrin

Figure 66 Peptide summary of valid search results for ICAT data (partial list)

Steps	Detailed Instructions	Comments
11 Examine the summary report.	**a** Check to see that your results are similar to those shown in Figure 66 on page 188. **b** Become familiar with the following headers: • **#** - table entry number • **Filename** - the name of the spectral file • **z** - peptide charge • **Score** - peptide database search score • **SPI (%)** - % scored peak intensity, which is the percentage of the ion current in the extracted MS/MS spectrum that is explained by the top database hit • **Spectrum Intensity** - intensity calculated from extracted ion chromatogram of precursor ion • **Sequence** - sequence from top database hit • **Cysteines** - identifies whether the precursor ion contained the light label, the heavy label, or no label (blank) • **L/H** - light to heavy ratio. If the Data Extractor cannot determine a charge for a peptide, it assumes a charge of +2 for determining the mass shifts for quantitation, and it looks for up to two modification sites (cysteines) in the peptide. When the actual charge is not +2, or when there are more than two modification sites in the peptide, the ratio is not calculated, and is reported as **n/c**. • **MH^+ Matched (Da)** - measured precursor ion MH^+ • **Accession #** - database accession number for the top hit • **Protein name** - name of the top database hit	• In the peptide mode, you may see a calculation discrepancy in cases where both members of a D_8/D_0 ICAT pair have been subjected to MS/MS. This is because the heavy- and light-labeled peptides often do not quite co-elute, so the EICs used for the calculations are slightly offset. (The EIC windows are determined by the Data Extractor setting for the time window for merging scans.) • You do not see this discrepancy in the protein mode, *provided* that both the D_8- and D_0-labeled precursor ions were subjected to MS/MS and that these results were of sufficient quality to be interpreted and included in the final results summary. In the protein modes, the peptide ratios are combined to calculate a ratio for the protein. In these modes, the ratio for the ICAT pair is recalculated directly from the EICs of each precursor, rather than the parallel EICs from the calculated *m/z* shifts from the precursor. • File names are in the format: Data_File_Name.aaaa.bbbb.c, where ◦ aaaa = first merged scan ◦ bbbb = last merged scan ◦ c = assigned precursor charge (0 means charge was ambiguous)

Exercise 7. Summarize valid results in protein mode

This exercise is similar to the previous one, except now you summarize valid results in protein mode.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is open, click the **Summary Settings** button near the top of the form to display the settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other pages.
2 Check that your **Data Directory** is set to **ICAT**.	• If you have just summarized ICAT results in peptide mode, your data directory should already be set correctly. If not, click the **Select...** button to select the **ICAT** folder.	
3 Set the **Mode** to **Protein Summary**.		Notice that the summary input form changes to correspond with the new display mode.
4 Set the **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Leave **Sort proteins by** set to the default of **Score**.		• The **Sort** parameters determine how the data are sorted in the results display. • When you process your own data, select the setting that is most helpful to you.
6 Set the **score** and **% SPI** filters to **> 0**.	**a** Set **Filter by protein score** to **> 0**. **b** For **Filter peptides by**, set both **Score** and **% SPI** to **> 0**.	When you set these thresholds to **> 0**, you are sure to display all the valid results.

Steps	Detailed Instructions	Comments
7 Filter results to show only proteins for which ICAT ratios have been calculated.	• In the **Required AAs** box, select **C**.	The software can perform amino acid filtering at either the MS/MS Search level or the Protein/Peptide Summary level.
8 Set the **Review Fields**.	**a** Mark the **Protein MW** and **DEQ ratios** check boxes. **b** Keep the other default settings. (Make sure the **Intensity** check box is marked.) **c** Check to see that your form looks like that in Figure 67 on page 192. Note that changes from defaults are highlighted in yellow.	• These settings specify the data fields that are displayed. • **DEQ ratios** are the differential expression quantitation ratios — the ratios of the light/heavy pairs. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		
10 Examine the overall summary report.	• Check that your results are similar to those in Figure 68 on page 192.	Note the colored cells in the summary report. • The color-code indicates relative quantities. Dark red is highest, orange is intermediate, and yellow is lowest. • The **Distinct Peptides** column gives the number of peptides detected for each protein, • The **Mean Peptide Spectral Intensity** is an average of the intensities for all peptides detected for that protein. These intensities are calculated from extracted ion chromatograms from the precursor ions. • Note that the ICAT ratio for the protein is calculated as the mean of the values for the peptides. The results include the mean, standard deviation, and number of values that contribute to the mean.

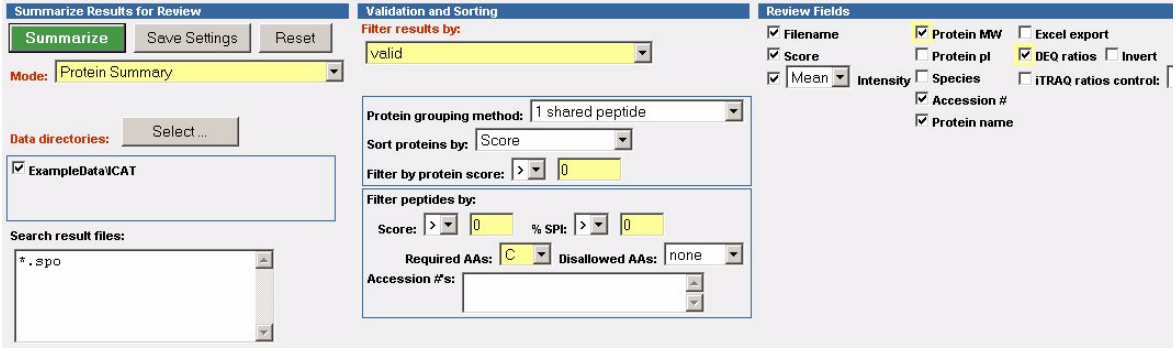
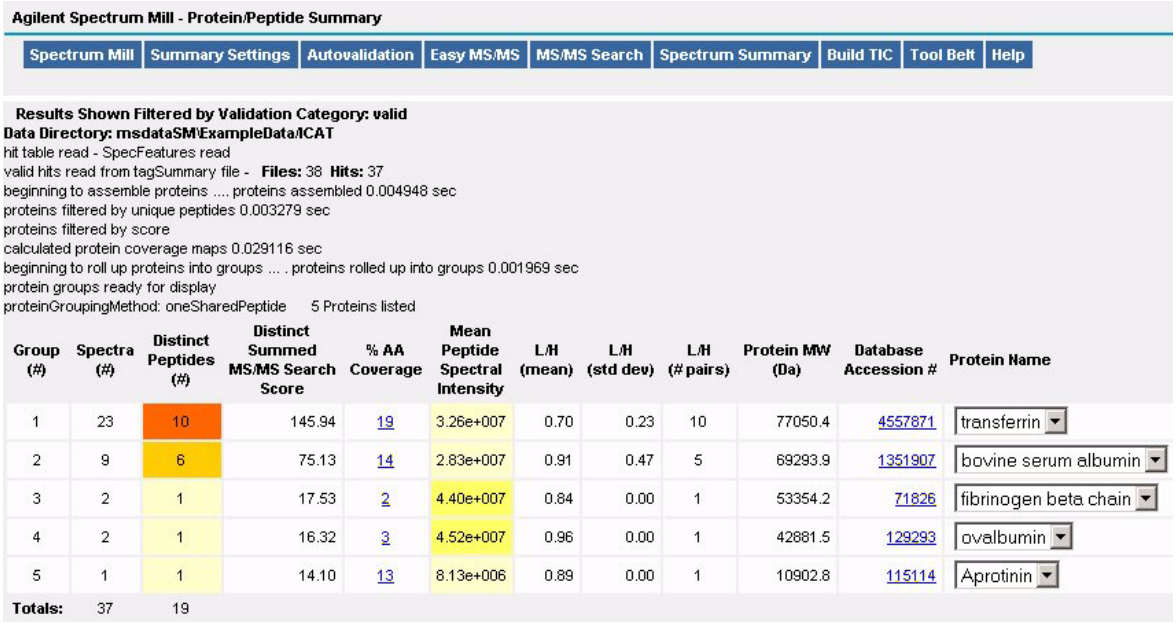


Figure 67 Settings to summarize valid ICAT results - Protein Summary mode



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\ICAT
hit table read - SpecFeatures read
valid hits read from tagSummary file - **Files:** 38 **Hits:** 37
beginning to assemble proteins .... proteins assembled 0.004948 sec
proteins filtered by unique peptides 0.003279 sec
proteins filtered by score
calculated protein coverage maps 0.029116 sec
beginning to roll up proteins into groups ... . proteins rolled up into groups 0.001969 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 5 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	L/H (mean)	L/H (std dev)	L/H (# pairs)	Protein MW (Da)	Database Accession #	Protein Name
1	23	10	145.94	19	3.26e+007	0.70	0.23	10	77050.4	4557871	transferrin
2	9	6	75.13	14	2.83e+007	0.91	0.47	5	69293.9	1351907	bovine serum albumin
3	2	1	17.53	2	4.40e+007	0.84	0.00	1	53354.2	71826	fibrinogen beta chain
4	2	1	16.32	3	4.52e+007	0.96	0.00	1	42881.5	129293	ovalbumin
5	1	1	14.10	13	8.13e+006	0.89	0.00	1	10902.8	115114	Aprotinin
Totals:	37	19									

Figure 68 Protein summary of valid search results for ICAT data

7
Processing iTRAQ Data

Exercise 1. (Optional) Transfer iTRAQ data file to Spectrum Mill server 194
Exercise 2. (Optional) Extract the spectra 196
Exercise 3. Search a database 201
Exercise 4. Autovalidate the hiqhest-quality results 206
Exercise 5. Create a file of iTRAQ correction factors 208
Exercise 6. Apply the iTRAQ correction factors to your sample 210
Exercise 7. Summarize valid results in peptide mode 212
Exercise 8. Summarize valid results in protein mode 216

This chapter shows how to process an iTRAQ data file. The iTRAQ reagents are a set of four isobaric reagents that react at the N-terminus and lysines. The reagents produce labeled peptides that are identical in mass and produce one signal in the MS spectrum, but that fragment in MS/MS mode to produce four unique signature ions. These low-mass signature ions allow you to simultaneously quantitate up to four different samples. In this exercise, you learn how to extract and search the spectra from an iTRAQ analysis, and how to perform the quantitation.

This chapter assumes that the Spectrum Mill workbench has already been installed on your server, the **SwissProt** database has been downloaded and indexed, and the server is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

Exercise 1. (Optional) Transfer iTRAQ data file to Spectrum Mill server

In this exercise, you move the iTRAQ data into the appropriate directory on the Spectrum Mill server.

NOTE The iTRAQ example data may have been installed on the server at the time the Spectrum Mill workbench was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

Steps	Detailed Instructions	Comments
1 Copy the example data directory, **X:\ExampleData\msdataSM\ExampleData\iTRAQ**, from the *Spectrum Mill Example Data CD* onto your client PC.	• Copy the **iTRAQ** directory from the *Spectrum Mill Example Data CD* to any directory on your client PC.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • This exercise simulates the likely laboratory scenario of file transfer from a client or instrument PC to the server; you may transfer files directly from the CD to the server if you prefer.
2 Create the **msdataSM\ExampleData\iTRAQ** folder on the Spectrum Mill server.	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In **msdataSM**, create **ExampleData\iTRAQ**.	• If you don't know how to find your Spectrum Mill file system, ask the person who installed your software. • **Do not include spaces or parentheses in your directory names.** • When you process your own samples, remember to set up a separate folder for each sample. Each folder should contain all sample fractions. • You may create up to ten folders between **msdataSM** and your data files, but shorter path lengths reduce memory usage, especially for large data sets.

Steps	Detailed Instructions	Comments
3 Copy the example files to the new directory.	**a** Copy or move the contents of the **iTRAQ** directory from your client PC to the new folder on the server PC. **b** Be sure to include both the **24prot_114_116.wiff** file and the **cpick_in** subdirectory.	• For convenience, these data are provided pre-extracted. The **cpick_in** subdirectory contains the extracted spectral files. • When you process your own data, remove any spaces or parentheses in the data file names. For example, change the file **my sample.d** to **my_sample.d**.
4 Make sure you have both read and write permissions for the data folders you just created.	**a** Right-click the **iTRAQ** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked, and apply the changes to all subfolders and files. **c** Click the **Security** tab. **d** Make sure all user groups have full permissions. **e** Click **OK**.	If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors.

Exercise 2. (Optional) Extract the spectra

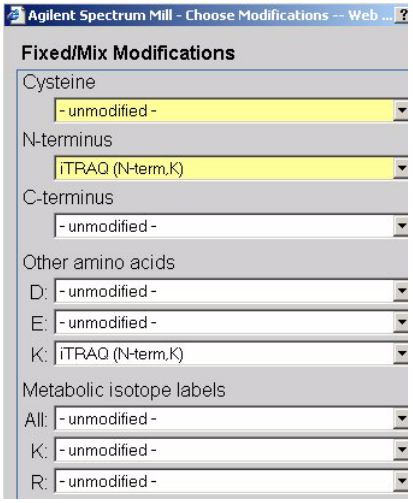
The Spectrum Mill Data Extractor preprocesses raw data files to extract high-quality spectra for database searches. For details about the processing that the Data Extractor accomplishes, see the introductory paragraph on page 16.

NOTE Since the iTRAQ data are provided pre-extracted, this exercise is optional. You need the optional Extractor License for Applied Biosystems/MDS Sciex QSTAR Data to do this exercise.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link.	

Steps	Detailed Instructions	Comments

2 Select the **msdataSM\ ExampleData\iTRAQ** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **iTRAQ** to select that data file directory. **c** Make sure that the name of the **iTRAQ** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 69 on page 200.	• Directories are identified by different types of icons:
		  ◦ Plain folders indicate directories that do not have data files directly beneath them.
		  ◦ Folders with rectangles indicate data directories.
		  ◦ Folders with line spectra (bar graphs) indicate data files.
		• Try clicking the names of each type to see which turn **bold**, indicating that they are selectable.
		• The software remembers your data file selection when you go to other Spectrum Mill forms.
		• If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Choose the appropriate modifications.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **-unmodified-**. **c** Under **N-terminus**, select **iTRAQ**. The software automatically adds **iTRAQ** as a modification to lysine (K).  **d** Click **OK.** The names of the modifications appear as in Figure 69 on page 200.	• The software also includes an **iTRAQ-Partial mix** modification, which you select if you suspect incomplete labeling. • To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
4 Set remaining parameters as shown in Figure 69 on page 200.	**a** Keep the defaults whenever appropriate. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. You may also need to change the time range for **Merge scans with same precursor m/z** to be compatible with your chromatographic data. **c** Click a blue section divider bar to display help for that section of the form.	• You use the **Sequence tag length** to filter out noisy spectra. This is the longest sequence of amino acids that is represented by the fragments in a spectrum. • If you set this to **>1**, you ensure that all possible good spectra are extracted. You have opportunities to set more stringent requirements later when you perform the database search.

Steps	Detailed Instructions	Comments
5 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** area to the right of the Data Extractor form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

6 (Optional) View the **cpick_in** subdirectory to see the files Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\iTRAQ\cpick_in**.
	b Notice the new files created there. Each one represents an extracted spectrum from your raw data file.
		• File names are in the format: Data_File_Name.aaaa.bbbb.c.pkl, where
		  ◦ aaaa = first merged scan
		  ◦ bbbb = last merged scan
		  ◦ c = assigned precursor charge (0 means charge was ambiguous).
		• For this instrument type, the great majority of extracted spectra have assigned precursor charges.

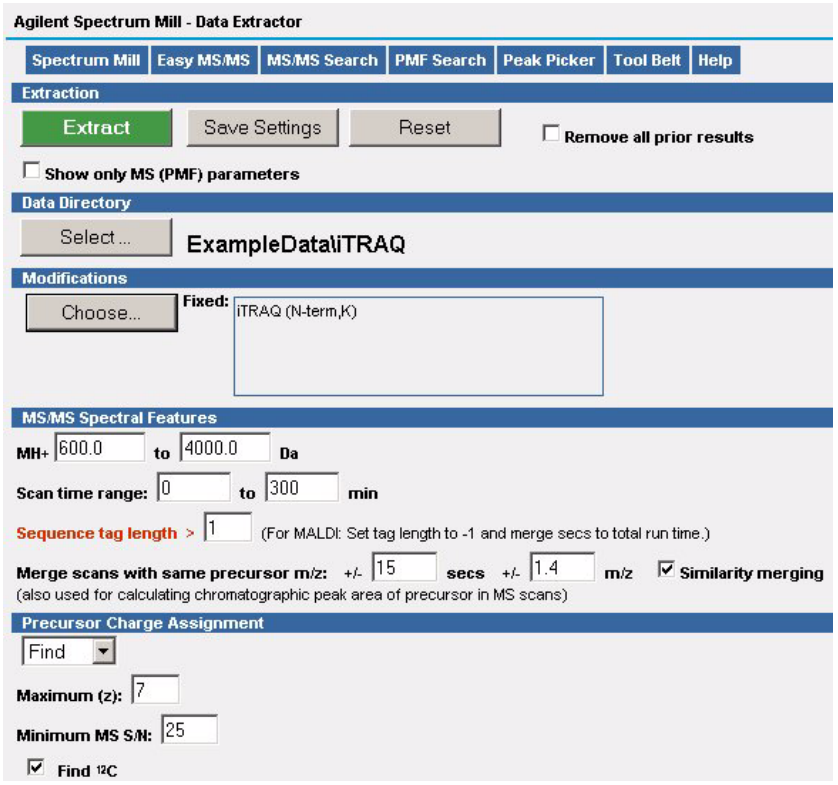
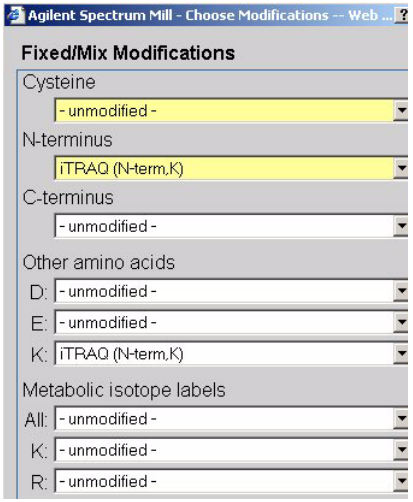


Figure 69 Settings for Data Extractor for iTRAQ data

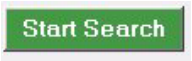
Exercise 3. Search a database

After you have extracted your spectra, you are ready to search each spectrum against a protein or DNA database. As you process data with the Spectrum Mill workbench, you may iterate through multiple rounds of database search and results validation, with the goal of identifying as many spectra as possible. For this initial search, you use identity mode, which means the spectra must be consistent with the database sequence. For later searches, you may use variable modifications or homology mode, which means spectra may show modifications relative to the database sequence.

Steps	Detailed Instructions	Comments
1 Make sure the public database **SwissProt** is available on your Spectrum Mill server.	**a** Ask your system administrator to install this database if that has not already been done. **b** Ask your system administrator to index this database.	To install and index a database, see the online help for **Protein Databases**. You can access this help from the Spectrum Mill home page.
2 Navigate to the MS/MS Search page.	• Do one of the following: ◦ From the Protein Databases page, click the **MS/MS Search** button. ◦ From the Spectrum Mill home page, click the link to MS/MS Search. 	Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Protein Databases page to the MS/MS Search page.
3 Check that your **Data Directory** is set to **iTRAQ**.	• If you have just performed the data extraction, your **iTRAQ** data directory should already be set correctly. If not, click the **Select...** button to select the **iTRAQ** folder.	

Steps	Detailed Instructions	Comments
4 Select the **SwissProt** database.	• For **Database** (under **Search Parameters**) select **SwissProt**.	If the database name does not appear, reload or refresh the MS/MS search form.
5 Choose the appropriate modifications.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **-unmodified-**. **c** Under **N-terminus**, select **iTRAQ**. The software automatically adds **iTRAQ** as a modification to lysine (K).  **d** Click **OK.** The names of the modifications appear as in Figure 70 on page 205.	• The modification is likely already set. • Three types of modifications are available for MS/MS Search: ◦ The **fixed** modifications are assumed to apply universally and are searched in a single search cycle. ◦ The **mix** modifications trigger cyclic MS/MS searches, where a different form of the modification is searched in each cycle. ◦ The **variable** modifications allow for both modified and unmodified forms within a peptide. All variable modifications are searched in each cycle. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
6 Set up the **Matching Tolerances**.	**a** Change the **Instrument** setting to **ESI QSTAR**. **b** Leave the other settings at default values.	

Steps	Detailed Instructions	Comments
7 Set up the **Search Mode**.	**a** Make sure that the check box for **Calculate reversed database scores** is marked. (This is the default setting.) **a** Mark the check box for **Dynamic peak thresholding**. **b** Verify that the **Search mode** is set to **Identity**.	• When you **Calculate reversed database scores**, you search against peptide sequences in their forward and inverted directions. If you obtain similar scores for both searches, you may have a false positive. • **Protein mobility scoring** is applicable only for ion trap data, and so is unavailable here. You do not use **Protein mobility scoring** with modifications such as iTRAQ that significantly alter the fragmentation patterns of peptides. • **Dynamic peak thresholding** is a scoring enhancement that enables identification of more low-abundance and short-chain peptides. For each extracted spectrum, the software calculates the search scores as the number of spectral peaks varies from n=4 up to the maximum set by the variable **peakLimitCount** in **instrument.txt**. It then displays the best score from the set.
8 Set other parameters as shown in Figure 70 on page 205.	**a** Keep the default settings. **b** Examine the items in red text carefully, since these are the ones you may need to change when you process your own samples. **c** Click a blue section divider bar to display help for that section of the form.	Changes from defaults are highlighted in yellow.

Steps	Detailed Instructions	Comments
9 Start the search.	**a** Click the **Start Search** button.  **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Scroll to the *top* of the **Results** area to see the message that indicates that the search is finished.	• MS/MS Search processes all spectral files in the directory. • Search time varies depending on the size of the database searched. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

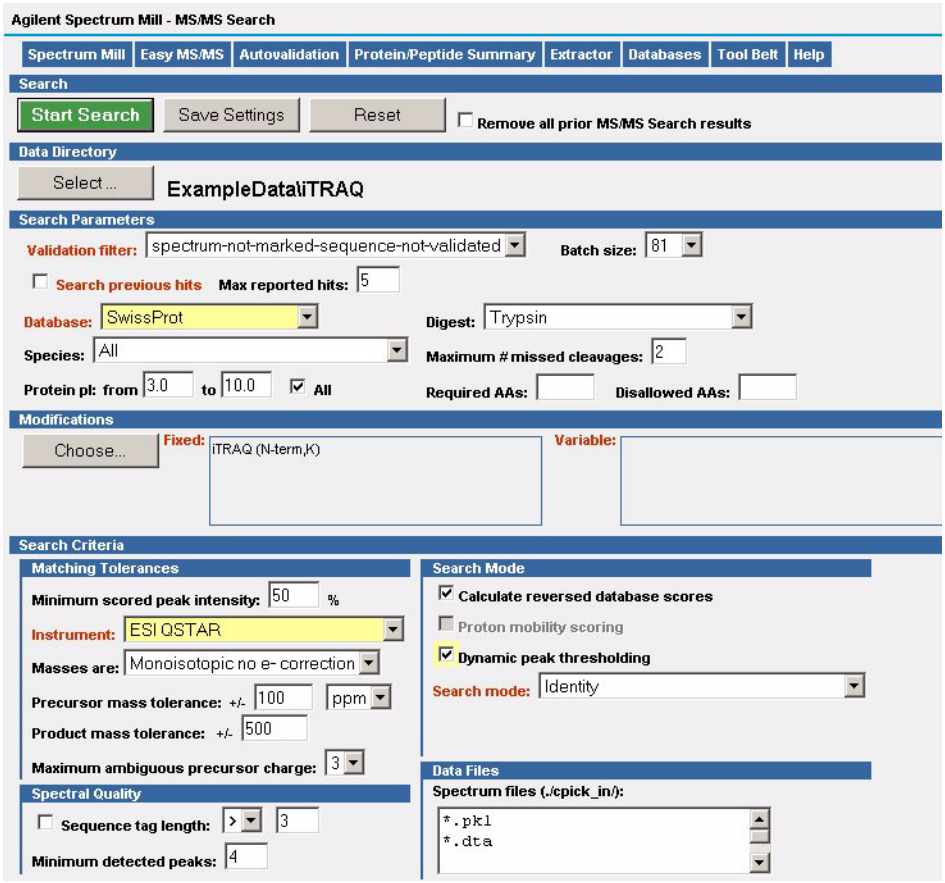


Figure 70 MS/MS Search settings

Exercise 4. Autovalidate the hiqhest-quality results

After you have completed a database search, you validate the good results. Validation means that you accept that the matches are correct.

The Spectrum Mill workbench provides a means to segregate search results that contain a valid interpretation of an MS/MS spectrum from those that do not. Results that are *not* validated can then be subjected to subsequent rounds of searches (against other databases or in variable modifications mode, for example). Results that *are* validated can be summarized in a results table.

The Spectrum Mill workbench provides two ways to validate results. One way uses the Autovalidation page, and is totally automated. You use this method only to validate the highest-scoring results — those that do not require manual review. The other method of validating uses the Protein/Peptide Summary page for manual review and validation. This exercise describes autovalidation.

Steps	Detailed Instructions	Comments

1 Navigate to the Autovalidation page.	• Navigate to this page from one of two other pages:
	  ◦ MS/MS Search
	  ◦ Protein/Peptide Summary
		You see the form shown in Figure 71 on page 207.

2 Check that your **Data Directory** is set to **iTRAQ**.	• If you have just performed the data extraction and MS/MS Search for this data set, your **iTRAQ** data directory should already be set correctly. If not, click the **Select...** button to select the **iTRAQ** folder.	
3 Validate first in the **Protein details** mode.	**a** For **Mode**, keep the default of **Protein details**. **b** Keep the default scoring presets, as shown in Figure 71. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by protein, and considers all the peptides that belong to a given protein. • Using the default scoring, individual peptides must have scores greater than 6 to 12 (depending on charge state), and the cumulative protein score must be greater than 20.

Steps	Detailed Instructions	Comments
4 Validate second in the **Peptide** mode.	**a** For **Mode**, select **Peptide**. **b** Keep the default scoring presets. **c** Click the **Validate Files** button. **d** Watch for a validation summary that lists the hits and spectra that have been validated.	• In this mode, the software summarizes results by peptide. Even if it finds only a single peptide for a protein, it validates the corresponding search results provided that the peptide score is high enough. • Using the default scoring, individual peptides must have scores greater than 11 to 15 (depending on charge state). This score threshold is higher than in the **Protein details** mode, where you have the additional assurance of knowing you have identified more than one peptide per protein.

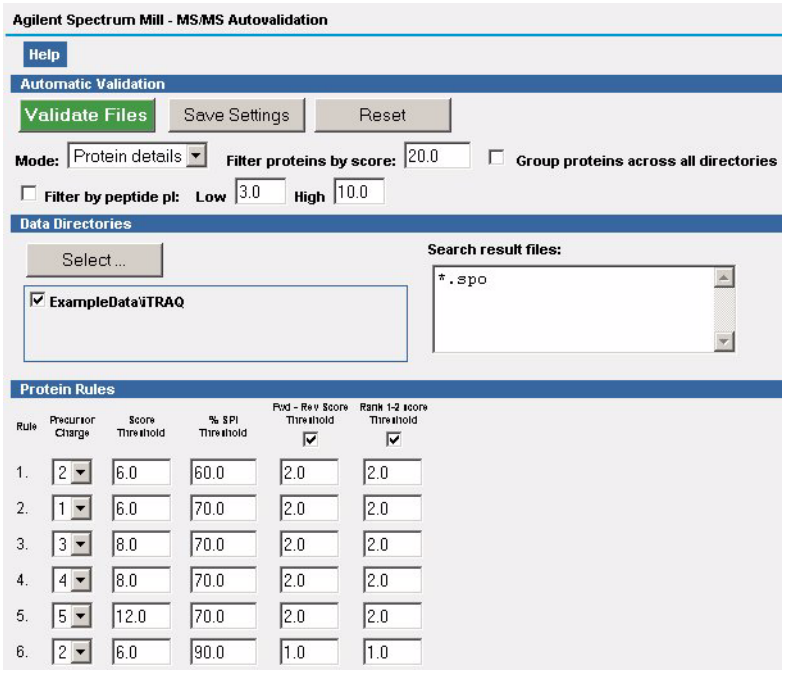


Figure 71 Autovalidation form, Protein details mode

Exercise 5. Create a file of iTRAQ correction factors

When you received your iTRAQ reagents, you received a certificate of analysis with iTRAQ correction factors. This exercise shows how to incorporate the correction factors into a file that you later apply to the iTRAQ calculations for your sample.

From a workflow standpoint, you could perform this exercise at any point prior to the next exercise. It does not need to follow autovalidation.

Steps	Detailed Instructions	Comments
1 Navigate to the Tool Belt page.	• From the Spectrum Mill home page, click the link to the Tool Belt page.	There are other ways to navigate to the Tool Belt page. Check the top of your current page for a **Tool Belt** button.
2 Click **Create iTRAQ correction factors**.		• You will create a text file that contains the iTRAQ correction factors. • You will later apply the correction factors to the data file used for the exercises in this chapter. • Once you have created the file with the correction factors, you can apply it to multiple samples.
3 Use your iTRAQ certificate of analysis to fill in the form. (Because you do not have the actual certificate of analysis for the sample in these exercises, see Figure 72 on page 209 for the numbers to type.)	**a** For **Batch**, type the kit lot number from your certificate of analysis. **b** For the rest of the form, type the software correction factors from your certificate of analysis.	When you process your own samples, type the correction factors from your certificate rather than the numbers shown in Figure 72.
4 Click the **Create** button.		• The software creates the file **iTRAQ.corrections** in the **msparams_mill** directory and **iTRAQbatches.js** in the **millhtml\sm_js** directory. • These files are appended every time you type correction factors for a new batch number.

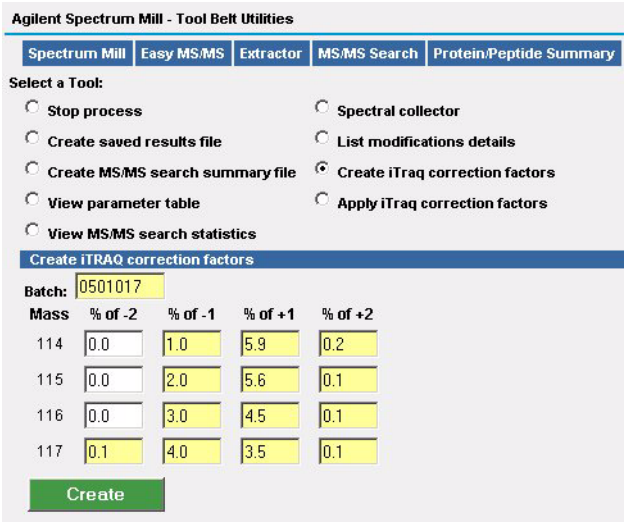


Figure 72 Example of form filled out with iTRAQ correction factors. When you process your own samples, type the correction factors from your iTRAQ certificate of analysis.

Exercise 6. Apply the iTRAQ correction factors to your sample

This exercise shows how to apply the iTRAQ correction factors to your data file so that the Spectrum Mill workbench can do all the final iTRAQ calculations. You do *not* need to export results to Excel to finalize the iTRAQ calculations.

Steps	Detailed Instructions	Comments
1 Navigate to the Tool Belt page.	• You are likely already there. If not, from the Spectrum Mill home page, click the link to the Tool Belt page.	There are other ways to navigate to the Tool Belt page. Check the top of your current page for a **Tool Belt** button.
2 Click **Apply iTRAQ correction factors**.		
3 Select the batch number.	• For **Batch** select the batch number that applies to your sample. For this exercise, select **0501017**.	
4 Set **Data Directory** to **iTRAQ**.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **iTRAQ** to select that data file directory. **c** Make sure that the name of the **iTRAQ** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 73 on page 211.	

Steps	Detailed Instructions	Comments
5 Click the **Apply** button.		• The software creates the file **iTRAQ.corrections.txt** in the **msdataSM** directory, in the subdirectory that contains your raw data file(s). • When the software performs iTRAQ calculations in Protein/Peptide Summary, it automatically applies the correction factors.

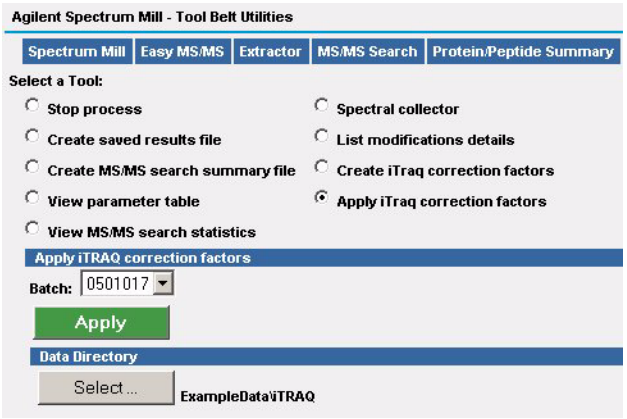


Figure 73 Application of iTRAQ correction factors to a specific data file directory

Exercise 7. Summarize valid results in peptide mode

Now, you summarize the results that you just automatically validated. In this exercise, you summarize in peptide mode, while in the next exercise, you summarize in one of the protein modes.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page, if you are not already there.	• From the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	There are links to this page from many other Spectrum Mill pages.
2 Check that your **Data Directory** is set to **iTRAQ**.	• If you have just performed autovalidation for this data set, your **iTRAQ** data directory should already be set correctly. If not, click the **Select...** button to select the **iTRAQ** folder.	
3 Keep the default **Mode** of **Peptide**.		• The Protein/Peptide Summary page offers a variety of results display modes. • These allow you to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set **Filter results by** to **valid**.		• The **Filter results by** setting selects from your data only the results that match your setting. • In this case, it selects only the results that match the **valid** setting (which was designated during autovalidation).
5 Leave the **Validation preset** set to the default of **none**.		• You use this setting when you want to summarize results rather than manually review and validate results.

Steps	Detailed Instructions	Comments
6 Set the **Sort** parameters.	• For **Sort peptides by**, keep the default of **Score**.	• The **Sort** parameters determine how the data are sorted in the results display. • In other modes, you may have an option for **Sort proteins by**. If you temporarily select the **Protein Summary Details** mode, you see this option. (Be sure to reset **Mode** to **Peptide** before you continue.)
7 Set the **Filter peptides by** parameters.	**a** Set the **Score** to **> 0**. **b** Set the **% SPI** to **> 0**.	By setting to zero, you ensure that all valid data is selected, regardless of score.
8 Set the **Review Fields**.	**a** Mark the check boxes for **iTRAQ intensities** and **iTRAQ ratios control**. **b** For **iTRAQ ratios control**, make sure that **114** is selected. **c** Mark the check boxes for both **N-terminus** and **Modifications**, because the reagents react at both the N-terminus and lysines. **d** For other settings, keep the defaults. **e** Check to see that your form looks like that in Figure 74 on page 214. Note that changes from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The **iTRAQ ratios control** is the reporter mass that the software uses as the control (denominator) for the ratio calculations. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
9 Click the **Summarize** button.		

Steps	Detailed Instructions	Comments
10 Examine the summary report.	**a** Check that your results are similar to those in Figure 75 on page 215. Since the public databases are constantly updated, your results may vary slightly. **b** Check the message about the iTRAQ correction factors, and make sure you applied the correct batch number. **c** Become familiar with the following headers that provide information about the iTRAQ labels: • **N-term** - shows the modification of the N-terminus • **Modifications** - shows modifications of amino acids other than C-terminal and N-terminal modifications • **iTRAQ** - columns show the intensities of the reporter ions, as well as their ratios	• File names are in the format: Data_File_Name.aaaa.bbbb.c, where • aaaa = first merged scan • bbbb = last merged scan • c = assigned precursor charge (0 means charge was ambiguous) • For this data type, the great majority of extracted spectra have assigned precursor charges.

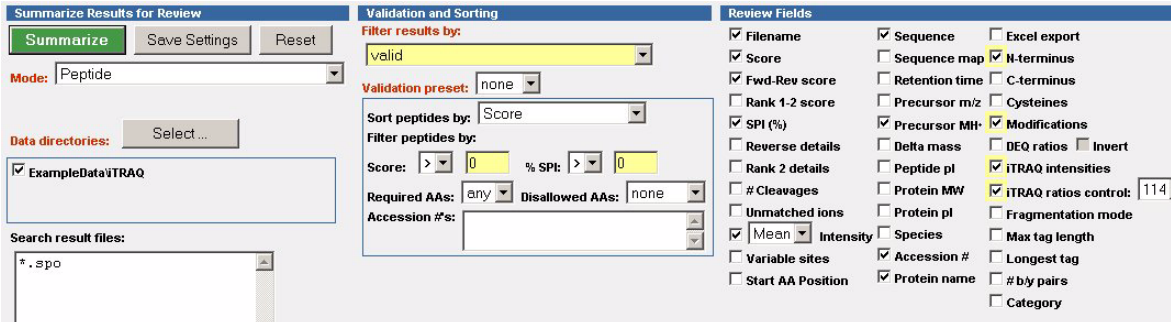
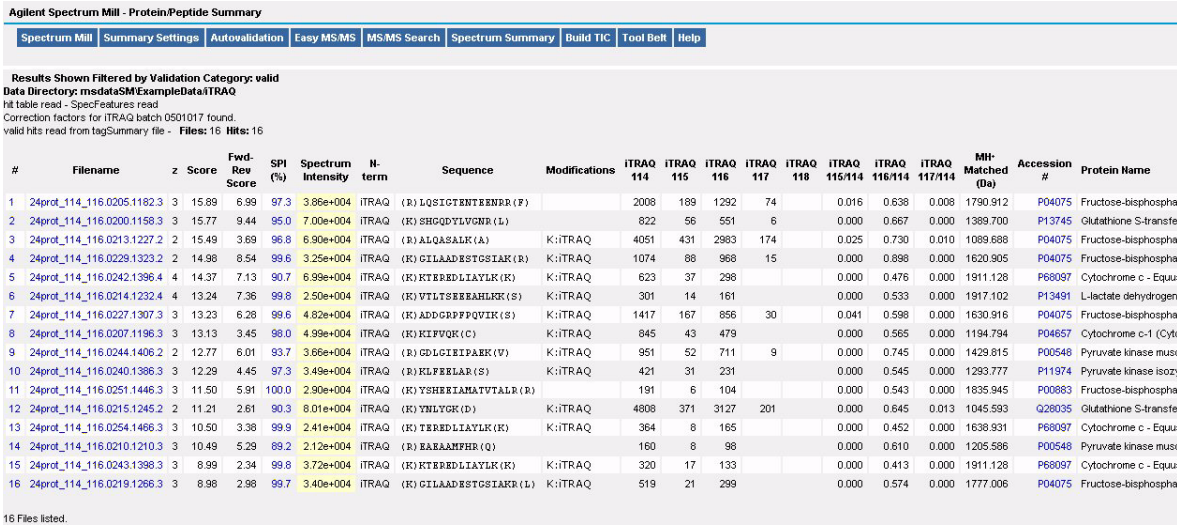


Figure 74 Settings to summarize valid iTRAQ results - peptide mode



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\iTRAQ
hit table read - SpecFeatures read
Correction factors for iTRAQ batch 0501017 found.
valid hits read from tagSummary file - **Files:** 16 **Hits:** 16

#	Filename	z	Score	Fwd-Rev Score	SPI (%)	Spectrum Intensity	N-term	Sequence	Modifications	iTRAQ 114	iTRAQ 115	iTRAQ 116	iTRAQ 117	iTRAQ 118	iTRAQ 115/114	iTRAQ 116/114	iTRAQ 117/114	MH+ Matched (Da)	Accession #	Protein Name
1	24prot_114_116.0205.1182.3	3	15.89	6.99	97.3	3.86e+004	iTRAQ	(R)LQSIGTENTEENRR(F)		2008	189	1292	74		0.016	0.638	0.008	1790.912	P04075	Fructose-bisphospha
2	24prot_114_116.0200.1158.3	3	15.77	9.44	95.0	7.00e+004	iTRAQ	(K)SHGQDYLVGNR(L)		822	56	551	6		0.000	0.667	0.000	1389.700	P13745	Glutathione S-transfe
3	24prot_114_116.0213.1227.2	2	15.49	3.69	96.8	6.90e+004	iTRAQ	(R)ALQASALK(A)	K:iTRAQ	4051	431	2983	174		0.025	0.730	0.010	1089.688	P04075	Fructose-bisphospha
4	24prot_114_116.0229.1323.2	2	14.98	8.54	99.6	3.25e+004	iTRAQ	(K)GILAADESTGSIAK(R)	K:iTRAQ	1074	88	968	15		0.000	0.898	0.000	1620.905	P04075	Fructose-bisphospha
5	24prot_114_116.0242.1396.4	4	14.37	7.13	90.7	6.99e+004	iTRAQ	(K)KTEREDLIAYLK(K)	K:iTRAQ	623	37	298			0.000	0.476	0.000	1911.128	P68097	Cytochrome c - Equu
6	24prot_114_116.0214.1232.4	4	13.24	7.36	99.8	2.50e+004	iTRAQ	(K)VTLTSEEEAHLKK(S)	K:iTRAQ	301	14	161			0.000	0.533	0.000	1917.102	P13491	L-lactate dehydrogen
7	24prot_114_116.0227.1307.3	3	13.23	6.28	99.6	4.82e+004	iTRAQ	(K)ADDGRPFPQVIK(S)	K:iTRAQ	1417	167	856	30		0.041	0.598	0.000	1630.916	P04075	Fructose-bisphospha
8	24prot_114_116.0207.1196.3	3	13.13	3.45	98.0	4.99e+004	iTRAQ	(K)KIFVQK(C)	K:iTRAQ	845	43	479			0.000	0.565	0.000	1194.794	P04657	Cytochrome c-1 (Cyt
9	24prot_114_116.0244.1406.2	2	12.77	6.01	93.7	3.66e+004	iTRAQ	(R)GDLGIEIPAEK(V)	K:iTRAQ	951	52	711	9		0.000	0.745	0.000	1429.815	P00548	Pyruvate kinase musc
10	24prot_114_116.0240.1386.3	3	12.29	4.45	97.3	3.49e+004	iTRAQ	(R)KLFEELAR(S)	K:iTRAQ	421	31	231			0.000	0.545	0.000	1293.777	P11974	Pyruvate kinase isozy
11	24prot_114_116.0251.1446.3	3	11.50	5.91	100.0	2.90e+004	iTRAQ	(K)YSHEEIAMATVTALR(R)		191	6	104			0.000	0.543	0.000	1835.945	P00883	Fructose-bisphospha
12	24prot_114_116.0215.1245.2	2	11.21	2.61	90.3	8.01e+004	iTRAQ	(K)YNLYGK(D)	K:iTRAQ	4808	371	3127	201		0.000	0.645	0.013	1045.593	Q28035	Glutathione S-transfe
13	24prot_114_116.0254.1466.3	3	10.50	3.38	99.9	2.41e+004	iTRAQ	(K)TEREDLIAYLK(K)	K:iTRAQ	364	8	165			0.000	0.452	0.000	1638.931	P68097	Cytochrome c - Equu
14	24prot_114_116.0210.1210.3	3	10.49	5.29	89.2	2.12e+004	iTRAQ	(R)EAEAAMFHR(Q)		160	8	98			0.000	0.610	0.000	1205.586	P00548	Pyruvate kinase musc
15	24prot_114_116.0243.1398.3	3	8.99	2.34	99.8	3.72e+004	iTRAQ	(K)KTEREDLIAYLK(K)	K:iTRAQ	320	17	133			0.000	0.413	0.000	1911.128	P68097	Cytochrome c - Equu
16	24prot_114_116.0219.1266.3	3	8.98	2.98	99.7	3.40e+004	iTRAQ	(K)GILAADESTGSIAKR(L)	K:iTRAQ	519	21	299			0.000	0.574	0.000	1777.006	P04075	Fructose-bisphospha

16 Files listed.

Figure 75 Peptide summary of valid search results for iTRAQ data

Exercise 8. Summarize valid results in protein mode

This exercise is similar to the previous one, except now you summarize valid results in protein mode.

Steps	Detailed Instructions	Comments
1 Navigate to the Protein/Peptide Summary page.	• If the Protein/Peptide Summary page is already open, click the **Summary Settings** button near the top of the page to display settings at the bottom of the page. • If the page is not open, from the Spectrum Mill home page, click the link to the Protein/Peptide Summary page.	
2 Check that your **Data Directory** is set to **iTRAQ**.	• If you have just performed results review for this data set, your **iTRAQ** data directory should already be set correctly. If not, click the **Select...** button to select the **iTRAQ** folder.	
3 Set display **Mode** to **Protein Summary**.		• When you process your own samples, you may choose another display mode, depending on whether you want to organize results by proteins, peptides, or samples. • For more details, see the MS/MS Data Review and Validation chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide*.
4 Set the **Filter results by** to **valid**.		• **Filter results by** selects from your data only the results that match your setting. • In this case, it selects only the valid results.

Steps	Detailed Instructions	Comments
5 For **Sort proteins by**, keep the default of **Score**.		• The **Sort** parameters determine how the data are sorted in the results display. • When you process your own data, select the setting that is most helpful to you.
6 Set the **score** and **% SPI** filters to **> 0**.	**a** Set **Filter by protein score** to **> 0**. **b** For **Filter peptides by**, set both **Score** and **% SPI** to **> 0**.	When you set these filters to **> 0**, you are sure to display all the valid results.
7 Set the **Review Fields**.	**a** Mark the **Protein MW** check box. **b** Mark the **iTRAQ ratios control** check box and select **114** as the denominator for the ratio calculations **c** Keep the other default settings. **d** Check to see that your form looks like that in Figure 76 on page 218. Note that settings that are changed from defaults are highlighted in yellow.	• **Review Fields** determine what information you see in the final results summary. • The default settings are shown in the online help. To access the online help, click the **Help** button for the form.
8 Click the **Summarize** button.		
9 Examine the overall summary report.	**a** Check that your results are similar to those in Figure 77 on page 218. **b** Check the message about the iTRAQ correction factors, and make sure you applied the correct batch number. **c** Note the columns that show the means and standard deviations of iTRAQ ratios, as well as the number of iTRAQ pairs.	For the proteins, the software calculates the means and standard deviations of the iTRAQ ratios from the component peptides.

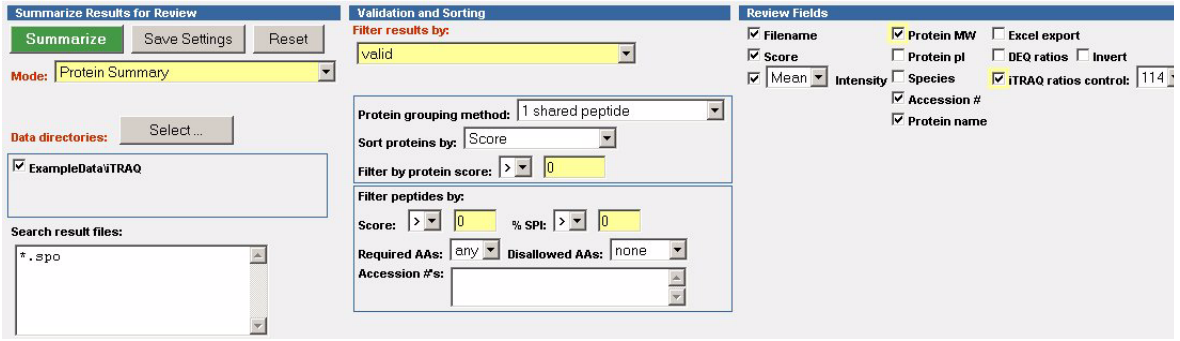
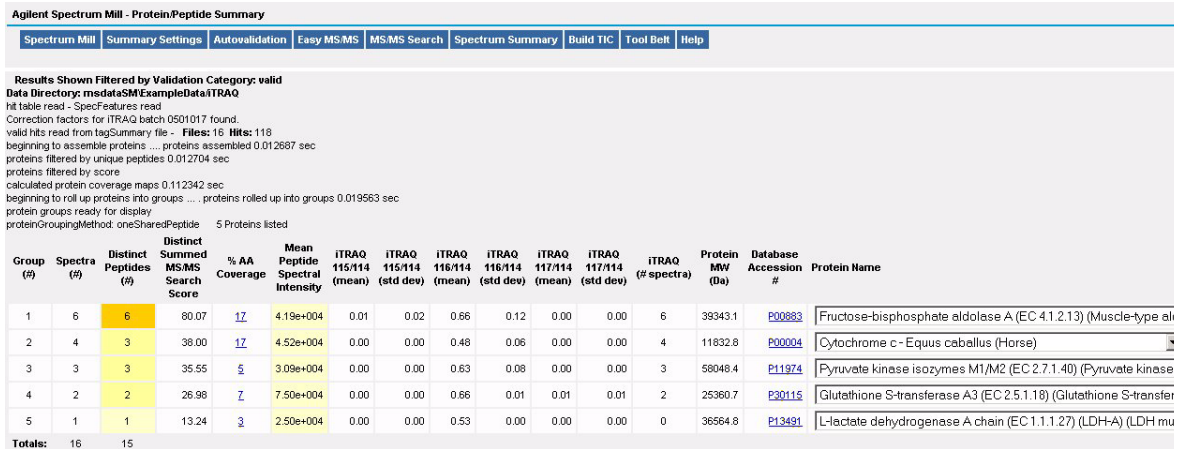


Figure 76 Settings to summarize valid iTRAQ results - Protein Summary mode



Agilent Spectrum Mill - Protein/Peptide Summary

Spectrum Mill | Summary Settings | Autovalidation | Easy MS/MS | MS/MS Search | Spectrum Summary | Build TIC | Tool Belt | Help

Results Shown Filtered by Validation Category: valid
Data Directory: msdataSM\ExampleData\iTRAQ
hit table read - SpecFeatures read
Correction factors for iTRAQ batch 0501017 found.
valid hits read from tagSummary file - **Files:** 16 **Hits:** 118
beginning to assemble proteins .... proteins assembled 0.012687 sec
proteins filtered by unique peptides 0.012704 sec
proteins filtered by score
calculated protein coverage maps 0.112342 sec
beginning to roll up proteins into groups .... proteins rolled up into groups 0.019563 sec
protein groups ready for display
proteinGroupingMethod: oneSharedPeptide 5 Proteins listed

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Spectral Intensity	iTRAQ 115/114 (mean)	iTRAQ 115/114 (std dev)	iTRAQ 116/114 (mean)	iTRAQ 116/114 (std dev)	iTRAQ 117/114 (mean)	iTRAQ 117/114 (std dev)	iTRAQ (# spectra)	Protein MW (Da)	Database Accession #	Protein Name
1	6	6	80.07	17	4.19e+004	0.01	0.02	0.66	0.12	0.00	0.00	6	39343.1	P00883	Fructose-bisphosphate aldolase A (EC 4.1.2.13) (Muscle-type al
2	4	3	38.00	17	4.52e+004	0.00	0.00	0.48	0.06	0.00	0.00	4	11832.8	P00004	Cytochrome c - Equus caballus (Horse)
3	3	3	35.55	5	3.09e+004	0.00	0.00	0.63	0.08	0.00	0.00	3	58048.4	P11974	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase
4	2	2	26.98	7	7.50e+004	0.00	0.00	0.66	0.01	0.01	0.01	2	25360.7	P30115	Glutathione S-transferase A3 (EC 2.5.1.18) (Glutathione S-transfer
5	1	1	13.24	3	2.50e+004	0.00	0.00	0.53	0.00	0.00	0.00	0	36564.8	P13491	L-lactate dehydrogenase A chain (EC 1.1.1.27) (LDH-A) (LDH mu
Totals:	16	15													

Figure 77 Protein summary of valid search results for iTRAQ data

8 First Steps for Processing MS-Only Data

Exercise 1. (Optional) Transfer data files to Spectrum Mill server 220
Exercise 2. Extract the LC-ESI-TOF data 222
Exercise 3. Create a species subset database 228
Exercise 4. Search the LC-ESI-TOF data 230
Exercise 5. Display PMF Search results for LC-ESI-TOF data 234
Exercise 6. Extract the AP-MALDI-TOF data 236
Exercise 7. Search the AP-MALDI-TOF data 240
Exercise 8. Display PMF Search results for AP-MALDI-TOF data 243
Exercise 9. Review AP-MALDI-TOF PMF Search results 246

This chapter provides familiarization exercises to process example LC-ESI-TOF and AP-MALDI-TOF data files with the Agilent Spectrum Mill MS Proteomics Workbench. The first data file is an LC-TOF analysis acquired with electrospray ionization (ESI). The second data file is a TOF analysis acquired with atmospheric pressure matrix-assisted laser desorption ionization (AP-MALDI). Both files contain MS-only data. Both examples are provided because the extraction and search settings are different for the two data types.

This chapter elaborates on the chapter about MS-only data in the companion *Spectrum Mill MS Proteomics Workbench Application Guide*. This familiarization chapter is intended as an introduction to the basic operation of the software; you are encouraged to customize these steps for your samples and studies.

This chapter assumes that the Spectrum Mill workbench has already been installed on your server, the SwissProt database has been downloaded and indexed, and the system is ready to go. If this has not yet been done, see the *Spectrum Mill MS Proteomics Workbench Installation Guide*.

Exercise 1. (Optional) Transfer data files to Spectrum Mill server

In this exercise, you copy AP-MALDI-TOF and LC-ESI-TOF data files to the Spectrum Mill server. One data file is the LC-ESI-TOF file that is used for Exercises 2-5. The other is the AP-MALDI-TOF file used for Exercises 6-9. After you copy the files to the server, you start the Spectrum Mill workbench.

NOTE The example data files may have been installed on the server at the time the Spectrum Mill workbench was installed. If so, you may either skip this exercise or (if you want the practice) delete the example data from the server and install it again with this exercise.

Steps	Detailed Instructions	Comments
1 Copy the example data file directories from the **X:\ExampleData\msdataSM\ExampleData\Agilent\TOF** directory on the *Spectrum Mill Example Data CD* onto your Spectrum Mill server.	**a** On your server, find the **SpectrumMill** folder. **b** In this folder, click to open the folder **msdataSM**. **c** In this folder, click to open the folder **ExampleData**. **d** In this folder, create a new folder and call it **Agilent**. **e** Within this folder, create another new folder and call it **TOF**. **f** Copy the **AP-MALDI** and **LC-ESI** folders from the *Spectrum Mill Example Data CD* to the **TOF** folder.	• You do this exercise outside the Spectrum Mill workbench. Use your normal file management utilities. • If you don't know how to find your Spectrum Mill file system, ask the person who installed your software. • **Do not include spaces or parentheses in your directory names.** • You may create up to ten folders between **msdataSM** and your data files.

Steps	Detailed Instructions	Comments
2 Make sure you have both read and write permissions for the data folders and files you just created.	**a** Right-click the **TOF** folder and select **Properties**. **b** Clear the **Read-only** check box if it is marked, and apply the changes to all subfolders and files. **c** Click the **Security** tab. **d** Make sure all user groups have full permissions. Again, apply changes to all subfolders and files. **e** Navigate to the **TOF\AP-MALDI** folder, right-click the **MALDI_TOF.wiff** file and clear the **Read-only** check box if it is marked. **f** Repeat step e for the **TOF\LC-ESI\ LC_TOF.wiff** file. **g** Click **OK**.	• If the Spectrum Mill workbench cannot write to the folders that contain your data files, you may encounter errors. • The software does not process *.wiff data files if they are marked **Read-only**.
3 Start the Spectrum Mill workbench on your PC.	• Double-click the desktop icon to launch the Spectrum Mill workbench.	• You may access the software from the server or a client PC. • If you have difficulty, see "Exercise 2. Access the Spectrum Mill server" on page 14.

Exercise 2. Extract the LC-ESI-TOF data

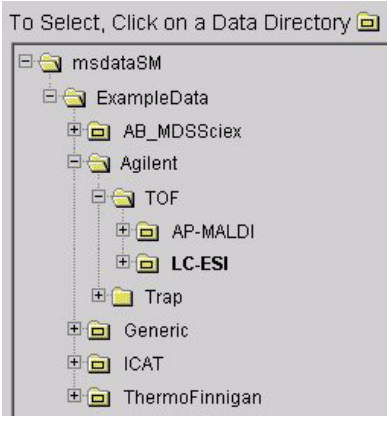
The MS-only data extractor performs different functions depending on the data to be extracted. For some data types, such as AP-MALDI-TOF data, the extractor averages and centroids the data from raw data files, but leaves signal-to-noise filtering for PMF Search. For other data types, such as LC-ESI-TOF data, the extractor performs the averaging, centroiding, and signal-to-noise filtering. You can also filter by signal-to-noise in PMF Search.

NOTE The Spectrum Mill workbench treats PDF-MALDI data the same as AP-MALDI data.

For LC-ESI-TOF data, the extractor does the following:

- Generates a base peak chromatogram (BPC)
- Generates a background peak list by signal-to-noise filtering (but not de-isotoping) the specified background range of the BPC
- Uses the **time segment** value to sample the BPC in “bins” of that width. Each “bin's” MS peak list is background subtracted, peak-picked (signal-to-noise filtering, de-isotoping) and charge assigned. The resulting peaks are merged with the prior running accumulation of peaks.
- After the file has been processed, the results are written to a *.MIC (Mass-Intensity-Charge) file.

This extractor filters the spectra so that all peaks are of sufficient quality to be submitted to PMF Search.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link. 	
2 Select the **msdataSM\ExampleData\Agilent\TOF\LC-ESI** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **LC-ESI** to select that data directory.  **c** Make sure that the name of the **LC-ESI** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 78 on page 226.	• Directories are identified by different types of icons: ◦ Plain folders indicate directories that do not have data files directly beneath them. ◦ Folders with rectangles indicate data directories. ◦ Folders with line spectra (bar graphs) indicate data files (which for Agilent files are also directories in themselves). • Try clicking the names of each type to see which turn **bold**, indicating that they are selectable. • The software remembers your data file selection when you go to other Spectrum Mill forms. • If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Set all parameters except those related to background subtraction.	**a** Mark the check box for **Show only MS (PMF) parameters**. This shortens the form for MS-only data. **b** Click the **Choose...** button near the middle of the form. Under the **Cysteine** heading, select **Carboxymethylation**. **c** Mark the check box for **Agilent ESI TOF data**.	• To view details about the modifications that are currently available on your server, click the **Details** button at the bottom right of the **Choose Modifications** dialog. • For more information about choosing modifications, see the online help. • Your system administrator can configure custom modifications.
4 Examine the data file in Analyst to determine an appropriate region to specify for background subtraction.	**a** Launch Analyst QS. **b** Open the data file **LC_TOF.wiff** in Analyst. **c** Note regions of the chromatogram that are suitably flat for background subtraction.	There are a number of regions that are suitable, but for purposes of this exercise, note the interval from 36 to 37 minutes.

5 Return to Spectrum Mill and set parameters related to background subtraction.	**a** Set **Extraction time range** to **0 to 37 min** (since there are no peaks after 37 min).
	b If necessary, click the **LC** option.
	c For **Background subtraction,** set to **Select range from: 36 to: 37 min**.
		• The form presents two ways to choose a time range for background subtraction:
		  ◦ **Select range from: _ to: _ min.** uses a range you specify
		  ◦ **Last _ minute of scan time range** uses the end of the data acquisition time
		• The time range for background subtraction does not need to be a subset of the **Extraction time range**. For example, you can select 0 to 1 minute for background subtraction, but then extract from 10 to 40 minutes.

Steps	Detailed Instructions	Comments
6 Set the **Time segment for averaging MS scans**.	• Keep the default of 10 secs.	• This value is used to sample the base peak chromatogram for mass spectral peaks in "bins" of the time segment width. The extraction time is proportional to this value. • If sequence coverage is not critical, use a segment value on the order of one-half a typical chromatographic peak width. • If coverage is critical, use a smaller value. • If the segment value you specify results in fewer than four scans, the extractor selects a value that gives at least four scans, and displays a warning in the **Results** window.
7 Make sure your form is set up as shown in Figure 78 on page 226.	**a** Note that with the exception of **Modifications**, settings that have been changed from defaults are highlighted in yellow. **b** Click a blue section divider bar to display help for that section of the form.	
8 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** window to the right of the Data Extractor form. See Figure 79 on page 227. **c** Look at the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions. • The number of extracted peaks varies from analysis to analysis. This number is displayed near the bottom of the **Results** window.

Steps	Detailed Instructions	Comments
9 (Optional) View the **fit_batch_in** subdirectory to see the file Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\Agilent\TOF\LC-ESI\fit_batch_in**. **b** Notice the new *.MIC file created there. This file shows mass, intensity, and charge for each extracted peak.	

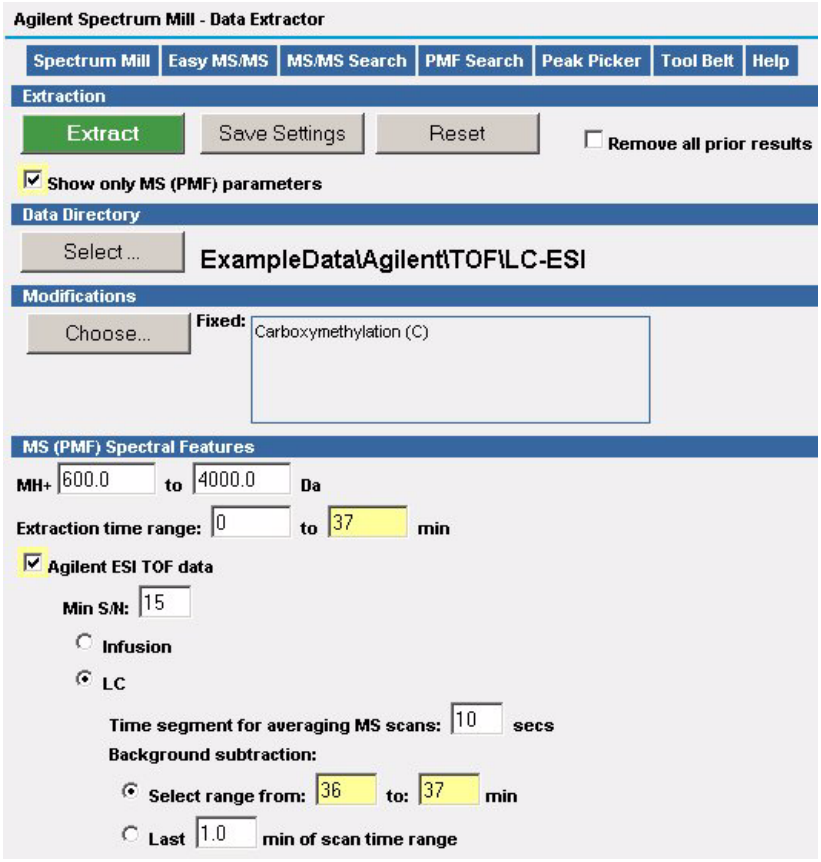


Figure 78 Data Extractor settings for LC-ESI-TOF data

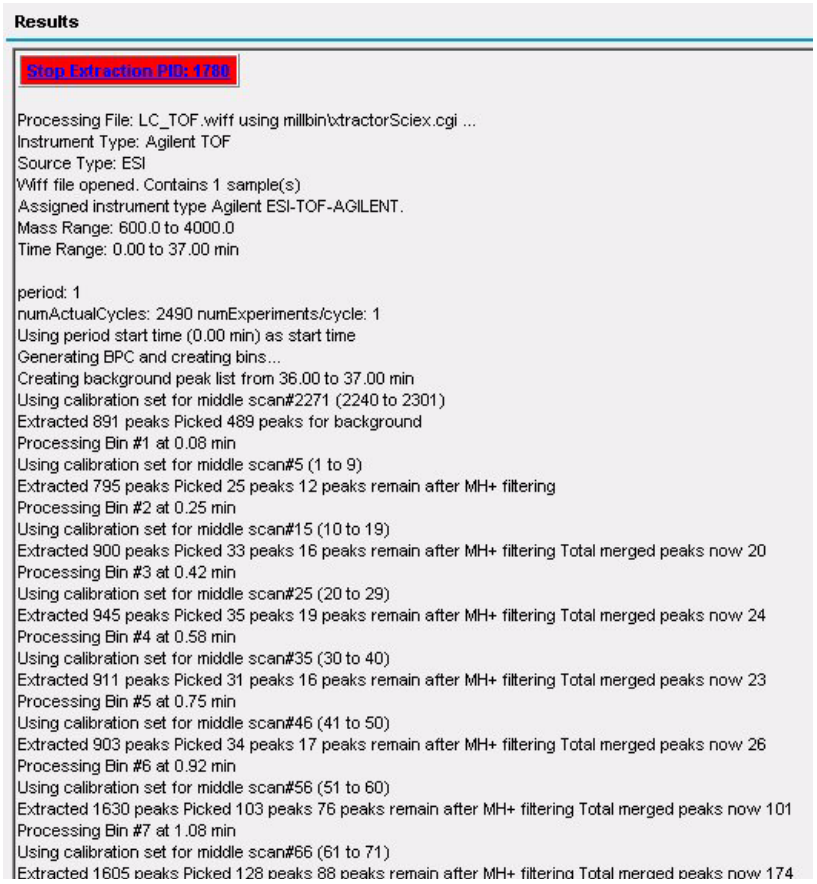
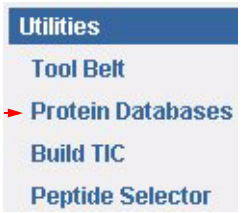

Results
Stop Extraction PID: 1780
Processing File: LC_TOF.wiff using millbin\xtractorSciex.cgi ...
Instrument Type: Agilent TOF
Source Type: ESI
Wiff file opened. Contains 1 sample(s)
Assigned instrument type Agilent ESI-TOF-AGILENT.
Mass Range: 600.0 to 4000.0
Time Range: 0.00 to 37.00 min
period: 1
numActualCycles: 2490 numExperiments/cycle: 1
Using period start time (0.00 min) as start time
Generating BPC and creating bins...
Creating background peak list from 36.00 to 37.00 min
Using calibration set for middle scan#2271 (2240 to 2301)
Extracted 891 peaks Picked 489 peaks for background
Processing Bin #1 at 0.08 min
Using calibration set for middle scan#5 (1 to 9)
Extracted 795 peaks Picked 25 peaks 12 peaks remain after MH+ filtering
Processing Bin #2 at 0.25 min
Using calibration set for middle scan#15 (10 to 19)
Extracted 900 peaks Picked 33 peaks 16 peaks remain after MH+ filtering Total merged peaks now 20
Processing Bin #3 at 0.42 min
Using calibration set for middle scan#25 (20 to 29)
Extracted 945 peaks Picked 35 peaks 19 peaks remain after MH+ filtering Total merged peaks now 24
Processing Bin #4 at 0.58 min
Using calibration set for middle scan#35 (30 to 40)
Extracted 911 peaks Picked 31 peaks 16 peaks remain after MH+ filtering Total merged peaks now 23
Processing Bin #5 at 0.75 min
Using calibration set for middle scan#46 (41 to 50)
Extracted 903 peaks Picked 34 peaks 17 peaks remain after MH+ filtering Total merged peaks now 26
Processing Bin #6 at 0.92 min
Using calibration set for middle scan#56 (51 to 60)
Extracted 1630 peaks Picked 103 peaks 76 peaks remain after MH+ filtering Total merged peaks now 101
Processing Bin #7 at 1.08 min
Using calibration set for middle scan#66 (61 to 71)
Extracted 1605 peaks Picked 128 peaks 88 peaks remain after MH+ filtering Total merged peaks now 174

Figure 79 Data Extractor running for LC-ESI-TOF data

Exercise 3. Create a species subset database

The Spectrum Mill workbench allows you to create subset databases based on species and/or molecular weight filters. You can search these subset databases much more quickly than a full database.

Since this example data file contain bovine proteins, you create a subset bovine (*bos taurus*) species database.

Steps	Detailed Instructions	Comments
1 Make sure the public database **SwissProt** is available on your Spectrum Mill server.	**a** Ask your system administrator to install this database if that has not already been done. **b** Ask your system administrator to index this database.	To install and index a database, see the online help for Protein Databases. You can access this help from the Spectrum Mill home page.
2 Navigate to the Protein Databases page.	• From the Spectrum Mill home page, click the Protein Databases link. 	
3 Open the form to create a species subset database.	• On the Protein Databases page, click the **Create species subset database** option.	
4 Fill in the form.	• Check that your form looks like the one shown in Figure 80 on page 229.	
5 Start the subset creation.	**a** Click the **Create** button. **b** After a short wait, check the bottom of the page for text that indicates that the software is creating the subset database. See Figure 81 on page 229 as an example.	The number of database entries depends on when you downloaded the **SwissProt** database.

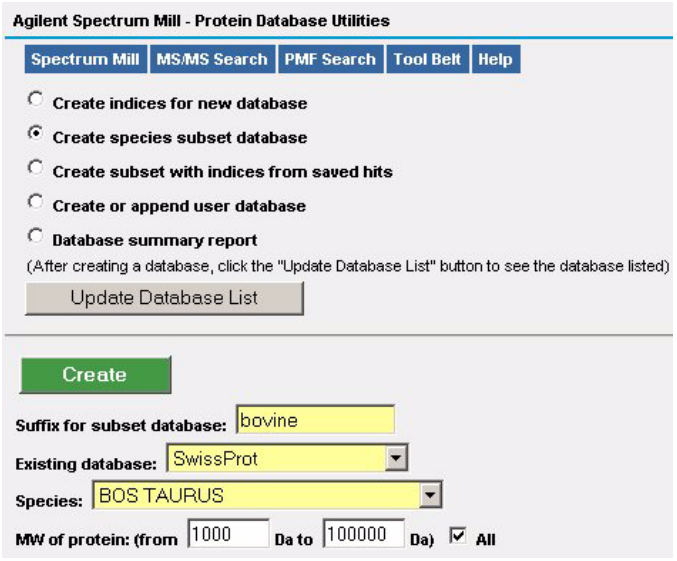

Agilent Spectrum Mill - Protein Database Utilities
Spectrum Mill
MS/MS Search
PMF Search
Tool Belt
Help
Create indices for new database
Create species subset database
Create subset with indices from saved hits
Create or append user database
Database summary report
(After creating a database, click the "Update Database List" button to see the database listed)
Update Database List
Create
Suffix for subset database: bovine
Existing database: SwissProt
Species: BOS TAURUS
MW of protein: (from 1000 Da to 100000 Da) All

Figure 80 Form to create species subset database

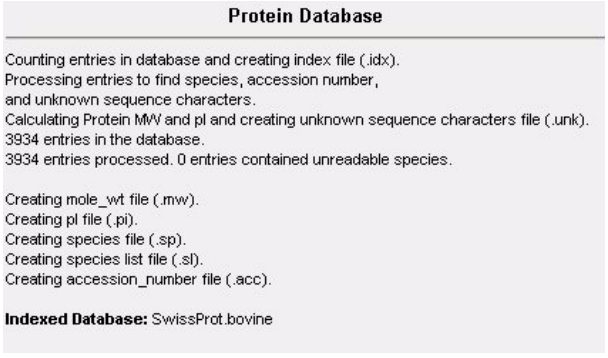

Protein Database
Counting entries in database and creating index file (.idx).
Processing entries to find species, accession number,
and unknown sequence characters.
Calculating Protein MW and pI and creating unknown sequence characters file (.unk).
3934 entries in the database.
3934 entries processed. 0 entries contained unreadable species.
Creating mole_wt file (.mw).
Creating pI file (.pi).
Creating species file (.sp).
Creating species list file (.sl).
Creating accession_number file (.acc).
Indexed Database: SwissProt.bovine

Figure 81 Creation of species subset database

Exercise 4. Search the LC-ESI-TOF data

After you have extracted your spectra, you are ready to search the peak list file against a protein or DNA database. If you do not achieve adequate identifications, you may repeat the search with another database.

Steps	Detailed Instructions	Comments
1 Navigate to the PMF Search page.	• Do one of the following: ◦ From the Data Extractor page, click the **PMF Search** button. ◦ From the Spectrum Mill home page, click the link to PMF Search. 	Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Data Extractor page to the PMF Search page.
2 Enable mixture scoring.	• Since this sample is known to contain a protein mixture, mark the check box for **Mixture scoring**.	• **Mixture scoring** assigns probability scores to potential mixtures and color-codes the mass spectrum to show peaks from each component. • You would mark the adjacent check box for **Remove all prior PMF results** if you executed a search with incorrect settings.
3 Check that your **Data Directory** is set to **LC-ESI**.	• If you have just performed the data extraction, your **LC-ESI** data directory should already be set correctly. If not, click the **Select...** button to select the **LC-ESI** folder.	

Steps	Detailed Instructions	Comments
4 Set the **Search Parameters**.	• For **Database**, select **SwissProt.bovine.**	If the database name does not appear, see your system administrator.
5 Set **Modifications**.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **Carboxymethylation**. **c** For the variable modifications, keep the defaults, as shown in Figure 82 on page 233.	
6 Set **Matching Tolerances**.	**a** Set **Instrument** to **Agilent ESI TOF**. **b** For **Minimum matched peptides,** retain the default of **5**. **c** For **Peptide mass tolerance,** retain the default of **+/- 5 ppm**.	Set the **Peptide mass tolerance** to a value that matches the mass accuracy you achieved for the analysis. • If you set it too wide, you risk false positives. • If you set it too narrow, you may exclude some peptides from the search.
7 Set **Spectral Features**.	**a** Mark the check box to **Override instrument defaults.** **b** Make sure **Max peaks** is set to the default of **500** and **Min S/N** is set to **2**.	• If you are concerned about ultimate coverage or you have a lot of proteins in your sample, set **Max peaks** to a larger number.
8 Set **Contaminant Masses**.	• Set **File** to **None**.	• For LC-ESI-TOF data, the extractor takes care of background subtraction, so there is no need to do it here. • To set up your own contaminant mass list file, click the **Contaminant Masses** blue bar to access the online help.
9 Confirm that your parameters are set as shown in Figure 82 on page 233.	**a** Note that with the exception of **Modifications**, settings that have been changed from defaults are highlighted in yellow. **b** Click a blue section divider bar to display help for that section of the form.	

Steps	Detailed Instructions	Comments
10 Start the search.	**a** Click the **Start Search** button. **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Look at the *top* of the **Results** area to see the message that indicates extraction is finished. 	• PMF Search processes all peak list files in the directory. • Search time varies depending on the size of the database searched. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

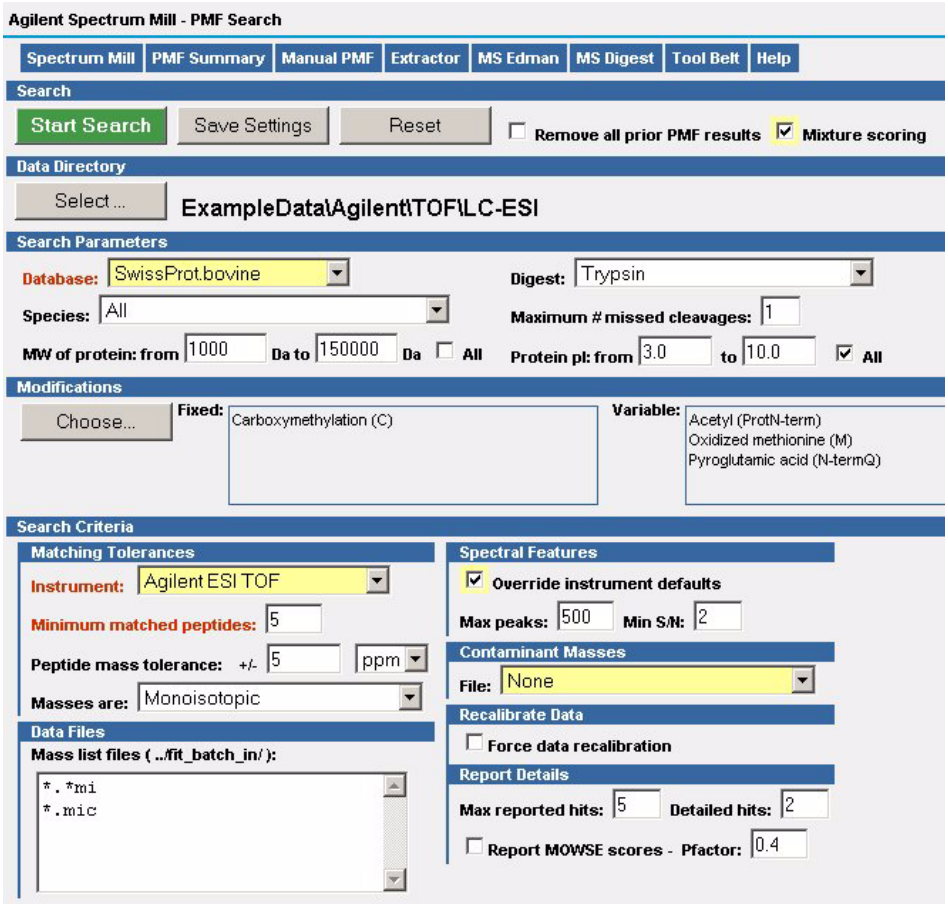
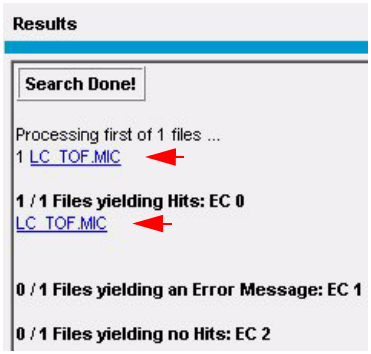


Figure 82 PMF Search settings for LC-ESI-TOF data

Exercise 5. Display PMF Search results for LC-ESI-TOF data

Now you display the PMF Search results so you can manually inspect them.

Steps	Detailed Instructions	Comments
1 Display the PMF Search results.	• In the **Results** area of the search page, click either of the links indicated by the arrows in the figure. 	• There are two ways to display results: ◦ Directly from the PMF Search page, as explained here ◦ From the PMF Summary page, as explained in "Exercise 8. Display PMF Search results for AP-MALDI-TOF data" on page 243.
2 Check that your results resemble those in Figure 83 on page 235.	**a** Note the score for mixture 1 (indicated with the arrow in the figure). **b** Note the three proteins that comprise the mixture (listed under the **Protein Name** header), as well as their individual scores. **c** Note the color coding in the spectrum Each mixture component is represented by a unique color.	Since the public databases are constantly updated, your results may vary.
3 View additional results.	**a** Scroll down to see additional mixture results. **b** Scroll all the way to the bottom to see details for the individual components.	For additional details about the display and review of PMF Search results, see "Exercise 8. Display PMF Search results for AP-MALDI-TOF data" and "Exercise 9. Review AP-MALDI-TOF PMF Search results".

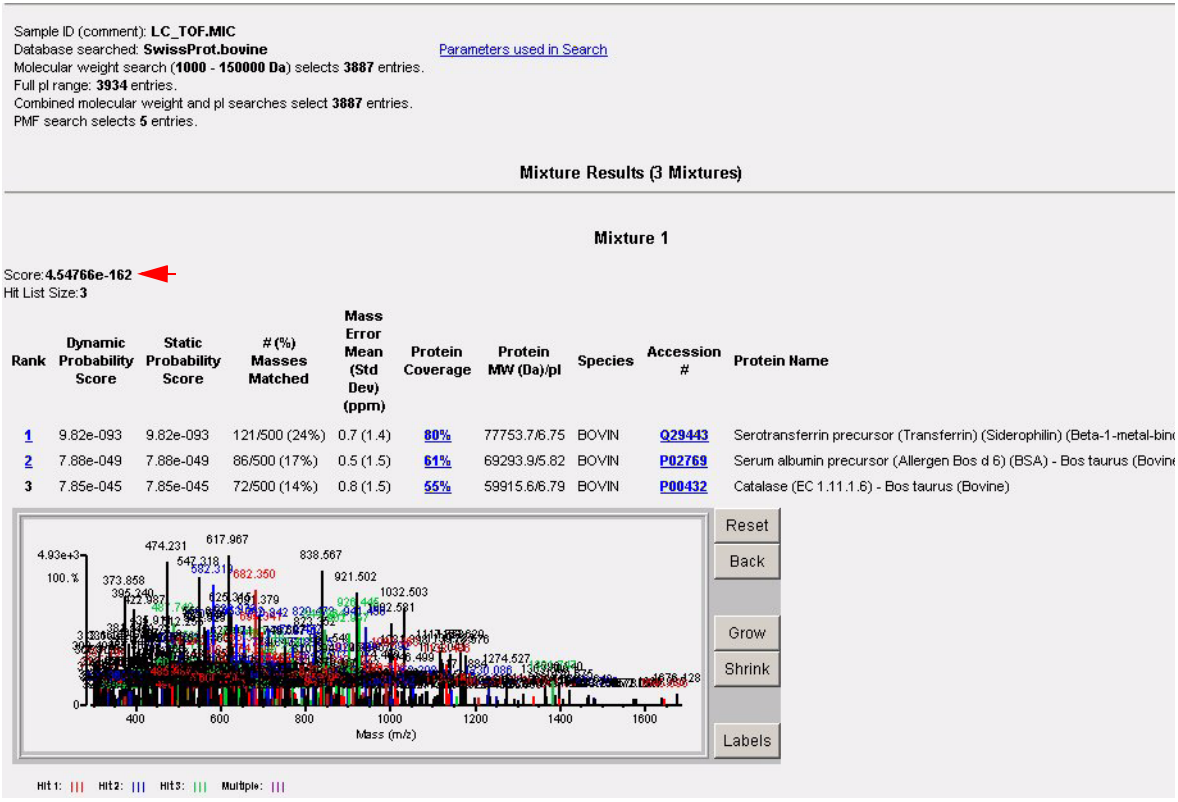


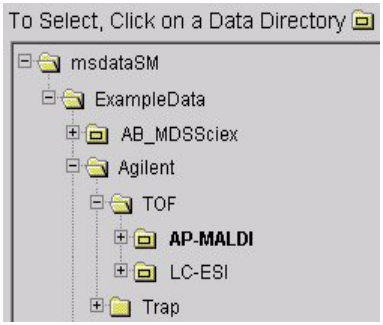
Figure 83 Mixture results for LC-ESI-TOF data

Exercise 6. Extract the AP-MALDI-TOF data

You process AP-MALDI-TOF data very similarly to LC-ESI-TOF data. The major differences lie in the data extraction and PMF Search. These two aspects are illustrated in detail here.

For AP-MALDI-TOF data, the extractor averages and centroids data from raw data files. In this case, the PMF Search de-isotopes, deconvolutes and filters the data by signal-to-noise.

NOTE The Spectrum Mill workbench treats PDF-MALDI data the same as AP-MALDI data.

Steps	Detailed Instructions	Comments
1 Navigate to the Data Extractor page.	• From the Spectrum Mill home page, click the Data Extractor link.	
2 Select the **msdataSM\ ExampleData\Agilent\TOF\ AP-MALDI** data directory.	**a** Click the **Select...** button near the middle of the form. **b** Expand the directory tree and click the word **AP-MALDI** to select that data directory.  **c** Make sure that the name of the **AP-MALDI** directory changes to a **bold** font. This indicates that it has been selected. **d** Click **OK.** The name of the data directory appears as in Figure 84 on page 239.	• Directories are identified by different types of icons: ◦ Plain folders indicate directories that do not have data files directly beneath them. ◦ Folders with rectangles indicate data directories. ◦ Folders with line spectra (bar graphs) indicate data files (which for Agilent files are also directories in themselves). • Try clicking the names of each type to see which turn **bold**, indicating that they are selectable. • The software remembers your data file selection when you go to other Spectrum Mill forms. • If you mark the **Make Default** check box in this dialog, the software remembers your data directory even after you close your web browser.

Steps	Detailed Instructions	Comments
3 Set parameters as shown in Figure 84 on page 239.	**a** Mark the check box for **Show only MS (PMF) parameters**. This shortens the form for MS-only data. **b** Click the **Choose...** button near the middle of the form. Under the **Cysteine** heading, select **Carboxymethylation**. **c** If necessary, clear the check box for **Agilent ESI TOF data**. **d** Note that with the exception of **Modifications**, settings that have been changed from defaults are highlighted in yellow. **e** Click a blue section divider bar to display help for that section of the form.	Unlike the LC-ESI-TOF extractor, the AP-MALDI-TOF extractor does not perform background subtraction using a background region from the data file. Later, PMF Search will subtract background masses set up in a contaminant mass list file.
4 Start the extraction.	**a** Click the **Extract** button. **b** View extraction progress in the **Results** area to the right of the Data Extractor form. **c** Look at the *top* of the **Results** area to see the message that indicates that extraction is finished.	• The Data Extractor processes all files in the directory. • Extraction time varies depending on the number and size of the files. • You can use your client PC for other tasks while the extraction takes place. • If you want to stop the extraction, click the red **Stop Extraction PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.
5 (Optional) View the **fit_batch_in** subdirectory to see the file Data Extractor has created.	**a** Navigate to the folder **SpectrumMill\msdataSM\ExampleData\Agilent\TOF\AP-MALDI\fit_batch_in**. **b** Notice the new *.XMI file created there.	

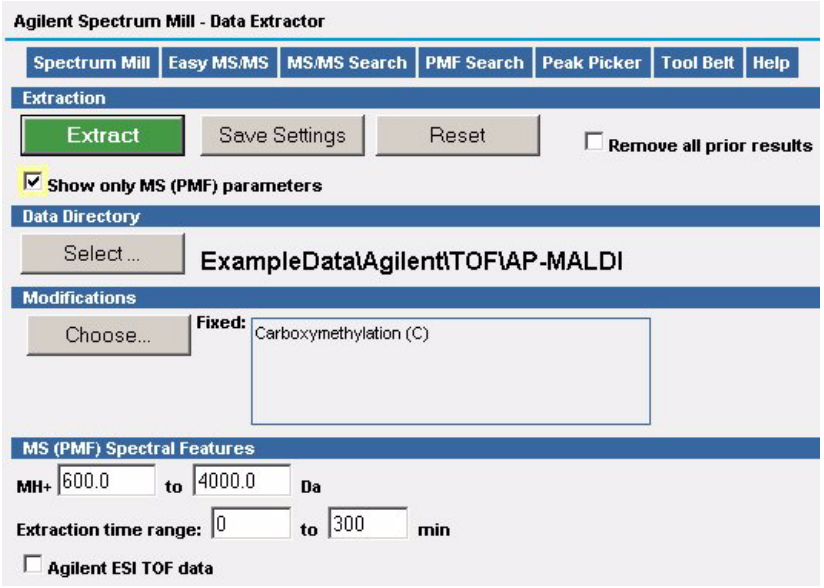


Figure 84 Data Extractor settings for AP-MALDI-TOF data

Exercise 7. Search the AP-MALDI-TOF data

After you have extracted your spectra, you are ready to search the peak list file against a protein or DNA database. If you do not achieve adequate identifications, you may repeat the search with another database.

Steps	Detailed Instructions	Comments

1 Navigate to the PMF Search page.	• Do one of the following:
	  ◦ From the Data Extractor page, click the **PMF Search** button.
	  ◦ From the Spectrum Mill home page, click the link to PMF Search.
		Since Spectrum Mill pages have buttons to take you to the next processing tasks, you can navigate directly from the Data Extractor page to the PMF Search page.

2 Establish correct settings for the check boxes near the top right-hand corner of the form.	**a** If necessary, clear the check box for **Remove all prior PMF results**. **b** If necessary, clear the check box for **Mixture scoring**. While the sample is a mixture, you clear the check box for purposes of this exercise. That way you can see differences from the results produced with the LC-ESI-TOF exercise, where this check box was marked.	• You would mark the check box for **Remove all prior PMF results** if you executed a search with incorrect settings and wanted to start over. • **Mixture scoring** assigns probability scores to potential mixtures and color-codes the mass spectrum to show peaks from each component.
3 Check that your **Data Directory** is set to **AP-MALDI**.	• If you have just performed the data extraction, your **AP-MALDI** data directory should already be set correctly. If not, click the **Select...** button to select the **AP-MALDI** folder.	
4 Set the **Search Parameters**.	• For **Database**, select **SwissProt.bovine.**	If the database name does not appear, see your system administrator.
5 Set **Modifications**.	**a** Click the **Choose...** button near the middle of the form. **b** Under the **Cysteine** heading, select **Carboxymethylation**. **c** For the variable modifications, keep the defaults, as shown in Figure 82 on page 233.	

Steps	Detailed Instructions	Comments
6 Set **Matching Tolerances**.	**a** For **Instrument**, retain the default setting of **Agilent AP-MALDI TOF**. **b** For **Minimum matched peptides**, retain the default of **5**. **c** For **Peptide mass tolerance**, retain the default of **+/-20 ppm**.	Set the **Peptide mass tolerance** to a value that matches the mass accuracy you achieved for the analysis. • If you set it too wide, you risk false positives. • If you set it too narrow, you may exclude some peptides from the search.
7 Set **Spectral Features**.	**a** Mark the check box to **Override instrument defaults.** **b** Make sure **Max peaks** is set to the default of **100**. **c** For **S/N**, keep the default of **2**.	The **Max peaks** setting of **100** is appropriate for this sample, which is composed of a three-protein-digest mixture.
8 Set **Contaminant Masses**.	• For **File**, retain the default of **Porcine Trypsin-Keratin**.	• This contaminant mass list file includes trypsin autolysis products and keratin. • To set up your own contaminant mass list file, click the **Contaminant Masses** blue bar to access the online help.
9 Confirm that your parameters are set as shown in Figure 85 on page 242.	**a** Note that with the exception of **Modifications**, settings that have been changed from defaults are highlighted in yellow. **b** Click a blue section divider bar to display help for that section of the form.	
10 Start the search.	**a** Click the **Start Search** button. **b** View search progress in the **Results** area to the right of the MS/MS Search form. **c** Look at the *top* of the **Results** area to see the message that indicates extraction is finished. 	• PMF Search processes all peak list files in the directory. • Search time varies depending on the size of the database searched. • You can use your client PC for other tasks while the search takes place. • If you want to stop the search, click the red **Stop Search PID: xxx** link at the top of the **Results** section. Then see the Tool Belt chapter in the *Spectrum Mill MS Proteomics Workbench Application Guide* for further instructions.

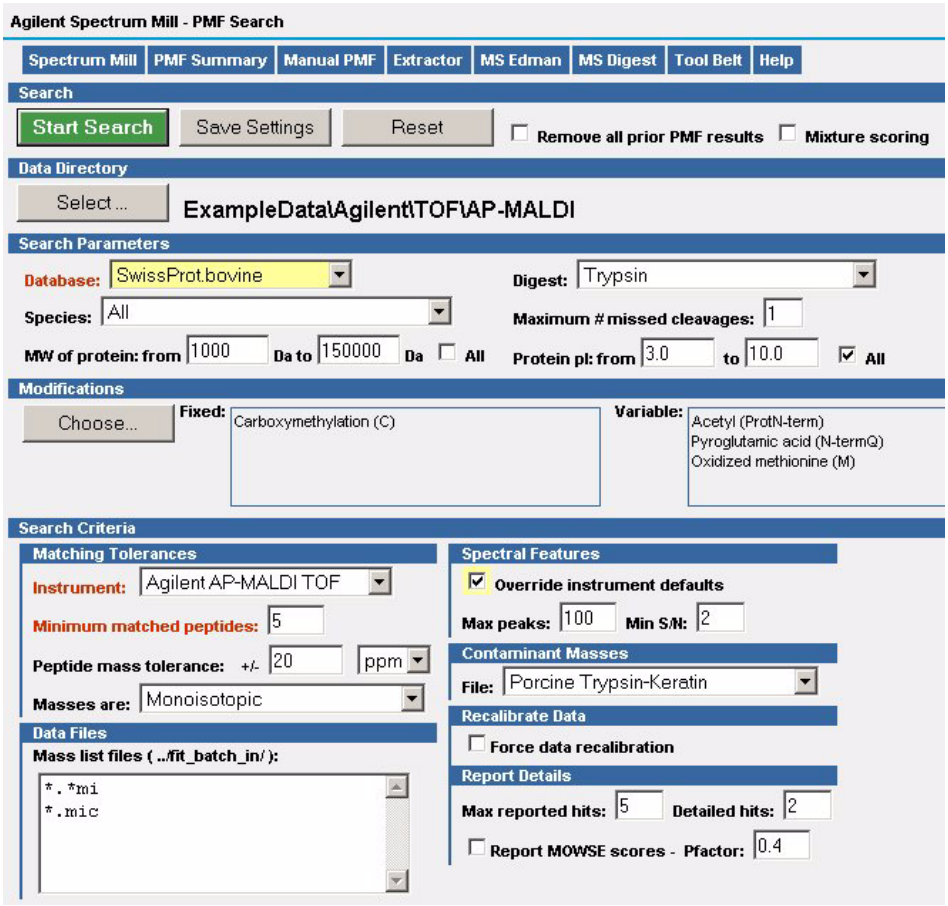


Figure 85 PMF Search settings for AP-MALDI-TOF data

Exercise 8. Display PMF Search results for AP-MALDI-TOF data

Now you display the PMF Search results so you can manually inspect them.

Steps	Detailed Instructions	Comments

1 Display the PMF Summary page.	• Do one of the following:
	  ◦ From the PMF Search page, click the **PMF Summary** button.
	  ◦ From the Spectrum Mill home page, click the link to the PMF Summary page.	

2 Check that your **Data Directory** is set to **AP-MALDI**.	• If you have just performed the PMF search for this data set, your **AP-MALDI** data directory should already be set correctly. If not, click the **Select...** button to select the **AP-MALDI** folder.	
3 Set the **Sorting** parameters.	• Keep the default settings.	• **Filter hits by score** selects from the data set only the members that exhibit the specified score criteria. • Since these scores represent probability of chance occurrence, smaller is better. • **Sort by** determines how the data are sorted in the results display.
4 Set the **Review Fields**.	**a** Keep the default settings. **b** Check to see that your form looks like that in Figure 86 on page 244.	**Review Fields** determine what information you see in the results summary.
5 Click the **Summarize** button.		

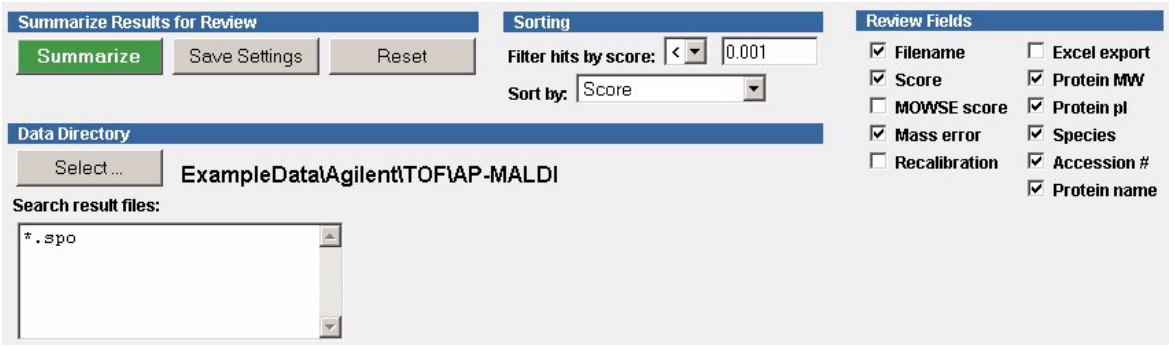


Figure 86 PMF Summary settings

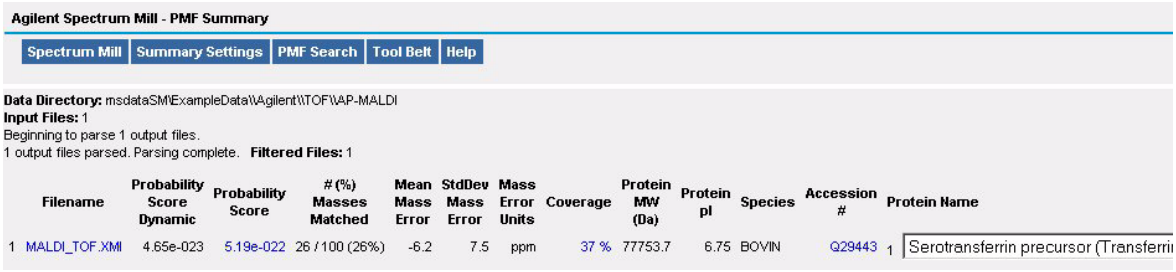


Figure 87 PMF Summary results for AP-MALDI-TOF data

Steps	Detailed Instructions	Comments
6 Examine the summary report.	**a** Check to see that your results are similar to those in Figure 87 on page 244. Note that since the public databases are constantly updated, your results may vary slightly. **b** Become familiar with the following headers: • **Filename** - the name of the (extracted) peak list file • **Probability Score Dynamic** - database search score based on the peptide mass tolerance determined by the Spectrum Mill software from the actual data • **Probability Score** - database search score based on the **Peptide mass tolerance** you set in PMF Search • **# (%) Masses Matched** - the fraction of the peaks submitted that were matched against the top database search result • **Mean Mass Error** - the average mass error for the matched peptides • **StdDev Mass Error** - the standard deviation of the mass error for the matched peptides • **Mass Error Units** - units for the mean mass error • **Coverage** - the percentage of the protein sequence accounted for by the matching peptide masses in the experimental spectrum • **Protein MW (Da)** - the molecular weight of the protein • **Protein pI** - the isoelectric point of the protein • **Species** - the species for the identified protein • **Accession number** - database accession number for the top match • **Protein name** - name of the top database match	The mass error was smaller for the LC-ESI-TOF data because that data used internal calibration, while this data used external calibration.

Exercise 9. Review AP-MALDI-TOF PMF Search results

After you display PMF Search results, you manually review them. In this exercise, you examine the PMF Search results to determine how well they fit the peak list that was submitted.

Steps	Detailed Instructions	Comments
1 Display the top hit.	**a** In the summary report, click the entry under **Filename**. **b** Check the bottom of the browser window, where you see the experimental spectrum displayed in the Spectrum Viewer. **c** Check that your results resemble those in Figure 88 on page 247. Because the public databases are constantly updated, your results may vary.	The experimental spectrum shows assigned peaks in color and unassigned peaks in black.
2 Use the Spectrum Viewer to examine details of the PMF search interpretation and its correlation with the spectrum.	**a** Learn to use the Spectrum Viewer. Follow the steps in "Exercise 10. Examine results with the Spectrum Viewer" on page 44. **b** Use the Spectrum Viewer to evaluate the search results.	The Spectrum Viewer used for PMF Summary has only a subset of the features described in the exercise.

Steps	Detailed Instructions	Comments
3 View a list of potential matches.	**a** Scroll up to see a list of potential matches, as shown in Figure 89 on page 248. Again, your results may vary. **b** Resize the page section to see more of your results on the screen: • Rest your mouse over the split line. • When the pointer becomes a double-sided arrow, drag the arrow to move the split line.	• While the summary results in Figure 87 on page 244 show only the top database match, the more detailed results shown in Figure 89 list multiple matches for each spectrum. • If this were ESI-TOF data, there could be duplicate PMF matches from singly- and multiply-charged peptides. In this case, only one match would be scored, but all would be displayed. The scored match would be the one for which **MH^+ Equivalent** was closest to **MH^+ matched**. In the **Score Counted** column, scored results would show a value of **1**, while non-scored results would show a value of **0**.

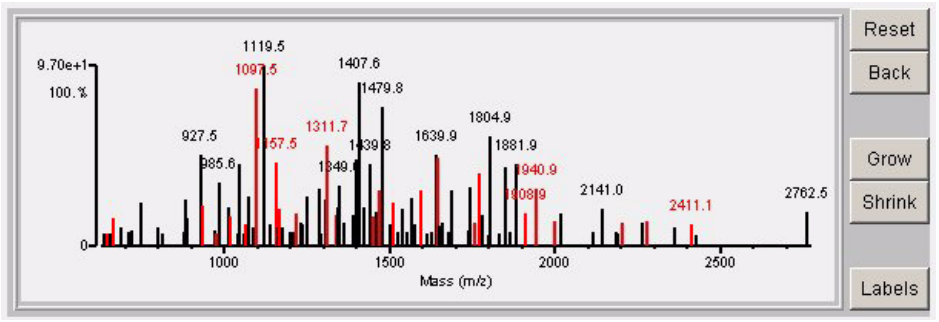
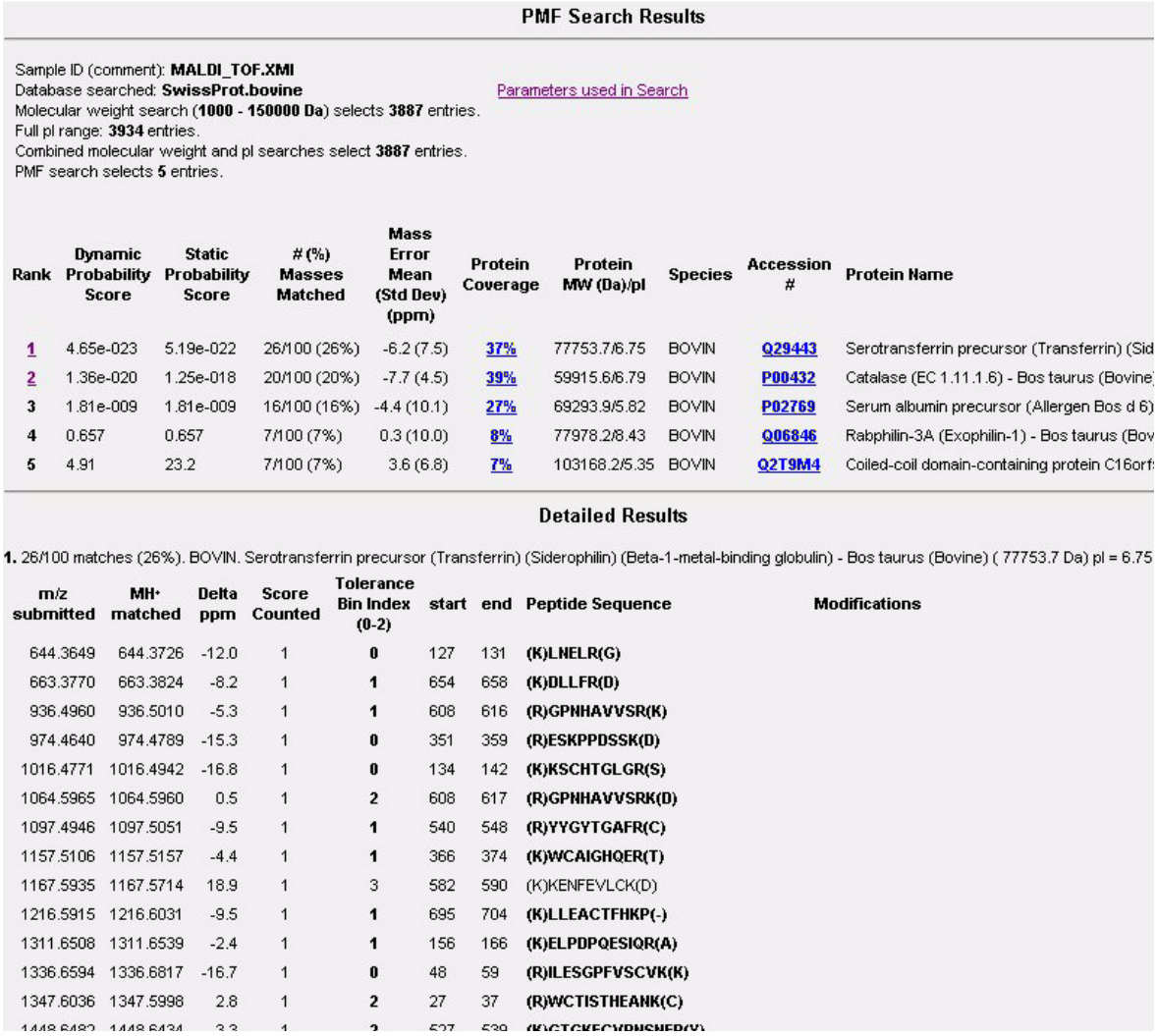


Figure 88 Spectrum Viewer



PMF Search Results

Sample ID (comment): **MALDI_TOF.XMI**
Database searched: **SwissProt.bovine** Parameters used in Search
Molecular weight search (**1000 - 150000 Da**) selects **3887** entries.
Full pI range: **3934** entries.
Combined molecular weight and pI searches select **3887** entries.
PMF search selects **5** entries.

Rank	Dynamic Probability Score	Static Probability Score	# (%) Masses Matched	Mass Error Mean (Std Dev) (ppm)	Protein Coverage	Protein MW (Da)/pI	Species	Accession #	Protein Name
1	4.65e-023	5.19e-022	26/100 (26%)	-6.2 (7.5)	37%	77753.7/6.75	BOVIN	Q29443	Serotransferrin precursor (Transferrin) (Sid
2	1.36e-020	1.25e-018	20/100 (20%)	-7.7 (4.5)	39%	59915.6/6.79	BOVIN	P00432	Catalase (EC 1.11.1.6) - Bos taurus (Bovine)
3	1.81e-009	1.81e-009	16/100 (16%)	-4.4 (10.1)	27%	69293.9/5.82	BOVIN	P02769	Serum albumin precursor (Allergen Bos d 6)
4	0.657	0.657	7/100 (7%)	0.3 (10.0)	8%	77978.2/8.43	BOVIN	Q06846	Rabphilin-3A (Exophilin-1) - Bos taurus (Bov
5	4.91	23.2	7/100 (7%)	3.6 (6.8)	7%	103168.2/5.35	BOVIN	Q2T9M4	Coiled-coil domain-containing protein C16orf

Detailed Results

1. 26/100 matches (26%). BOVIN. Serotransferrin precursor (Transferrin) (Siderophilin) (Beta-1-metal-binding globulin) - Bos taurus (Bovine) (77753.7 Da) pI = 6.75

m/z submitted	MH+ matched	Delta ppm	Score	Counted	Tolerance Bin Index (0-2)	start	end	Peptide Sequence	Modifications
644.3649	644.3726	-12.0	1		**0**	127	131	**(K)LNELR(G)**	
663.3770	663.3824	-8.2	1		**1**	654	658	**(K)DLLFR(D)**	
936.4960	936.5010	-5.3	1		**1**	608	616	**(R)GPNHAVVSR(K)**	
974.4640	974.4789	-15.3	1		**0**	351	359	**(R)ESKPPDSSK(D)**	
1016.4771	1016.4942	-16.8	1		**0**	134	142	**(K)KSCHTGLGR(S)**	
1064.5965	1064.5960	0.5	1		**2**	608	617	**(R)GPNHAVVSRK(D)**	
1097.4946	1097.5051	-9.5	1		**1**	540	548	**(R)YYGYTGAFR(C)**	
1157.5106	1157.5157	-4.4	1		**1**	366	374	**(K)WCAIGHQER(T)**	
1167.5935	1167.5714	18.9	1		3	582	590	(K)KENFEVLCK(D)	
1216.5915	1216.6031	-9.5	1		**1**	695	704	**(K)LLEACTFHKP(-)**	
1311.6508	1311.6539	-2.4	1		**1**	156	166	**(K)ELPDPQESIQR(A)**	
1336.6594	1336.6817	-16.7	1		**0**	48	59	**(R)ILESGPFVSCVK(K)**	
1347.6036	1347.5998	2.8	1		**2**	27	37	**(R)WCTISTHEANK(C)**	
1448.6482	1448.6434	3.3	1		**2**	527	539	**(K)GTGKECVPNSNEP(V)**	

Figure 89 Detailed PMF results for AP-MALDI-TOF data

Steps	Detailed Instructions	Comments
4 View the detailed results for the second hit.	**a** In the list shown in Figure 89, click the second entry under **Rank**. **b** Check that you see a list of *m/z* submitted, along with peptide sequences and other information. **c** Click the **Back** arrow on your browser window to go back to the previous screen.	• The hit list shown in Figure 89 displays only the number of hits that you specified in the **Max reported hits** box in the PMF Search form. • In the hit list shown in Figure 89, you can link to detailed results for only the number of hits that you specified in the **Detailed hits** box in the PMF Search form.
5 View a coverage map.	**a** Click a link under the **Protein Coverage** header to view amino acid coverage. **b** Close the coverage window.	In the coverage map: • Red = matched • Black = not matched • Blue = modified amino acid, e.g., cysteine • Green = consensus N-linked glycosylation site
6 View information in the database.	**a** Click a link under the **Accession #** header to the database information. **b** Close the accession number window.	

Steps	Detailed Instructions	Comments
7 (Optional) Generate a printout of the results.	**a** If you have not already done so, enable **Print background colors and images** in Internet Explorer. • Select **Tools > Internet Options...**. • Click the **Advanced** tab. • Mark the check box for **Print background colors and images**. • Click **OK**. **b** Do one of the following: • In Internet Explorer (IE) 6.0, select **File > Page Setup...** and set the page to landscape mode. Click **OK**. • In IE 7.0, click the arrow to the right of the printer icon, select **Page Setup...** and set the page to landscape mode. Click **OK**. **c** Click in the frame you wish to print. **d** Do one of the following: • In IE 6.0, select **File > Print Preview...**. • In IE 7.0, click the arrow to the right of the printer icon and select **Print Preview...**. **e** At the top of the page, select **Only the selected frame**. **f** Click the **Print...** button.	

www.agilent.com

In This Book

The *Familiarization Guide* presents exercises to process example data with the Spectrum Mill MS Proteomics Workbench. In this guide you learn:

- How to extract spectra from MS/MS and MS-only data sets
- How to perform protein database searches
- How to review and validate search results
- How to perform *de novo* sequencing
- How to process data from Agilent ion traps (CID and ETD), Agilent Q-TOF, and Agilent TOF
- How to process ICAT and iTRAQ data

© Agilent Technologies, Inc. 2007

Printed in USA
Fifth edition, June 2007

G2721-90032

Agilent Technologies